Sequence and structure of epoxide hydrolases: a systematic analysis

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

Institute of Technical Biochemistry, University of Stuttgart, Germany

Corresponding author:

Jürgen Pleiss, Ph.D.

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

Short title:

A systematic analysis of epoxide hydrolases

Keywords:

epoxide hydrolase, phylogenetic tree, superfamily, structure, sequence, annotation

Abstract

Epoxide hydrolases (EC 3.3.2.3) are ubiquitous enzymes which catalyze the hydrolysis of epoxides to the corresponding vicinal diols. Over 100 epoxide hydrolases (EH) have been identified or predicted, 3 structures are available. Although they catalyze the same chemical reaction, sequence similarity is low. To identify conserved regions, all EHs were aligned. Phylogenetic analysis identified 12 homologous families, which were grouped into 2 major superfamilies: the microsomal EH superfamily, which includes the homologous families of Mammalian, Insect, Fungal, and Bacterial EHs, and the cytosolic EH superfamily, which includes Mammalian, Plant, and Bacterial EHs. Bacterial EHs show a high sequence diversity. Based on structure comparison of 3 known structures from *Agrobacterium radiobacter* AD1 (cytosolic EH), *Aspergillus niger* (microsomal EH), and *Mus musculus* (cytosolic EH), and multisequence alignment and phylogenetic analysis of 95 EHs, the modular architecture of this enzyme family was analyzed. While core and cap domain are highly conserved, the structural differences between the EHs are restricted to only 2 loops: the NC-loop connecting the core and the cap and the cap-loop which is inserted into the cap domain. EHs were assigned to either of 3 clusters based on loop length. Using this classification, core and cap region of all EHs, NC-loops and cap-loops of 78% and 89% of all EHs, respectively, could be modeled. Representative models are available from the *Lipase Engineering Database*, http://www.led.uni-stuttgart.de.

Introduction

Epoxide hydrolases (EHs; EC 3.3.2.3) are a diverse group of functionally related enzymes, which catalyze the hydrolysis of epoxides to their vicinal diols and thus play a key role in detoxification [1]. EHs are cofactor independent and have been found in various organisms including plants, insects, mammals and bacteria [2-5]. EHs belong to the α/β hydrolase fold family [6] and consist of a core domain, a β-sheet packed between 2 layers of α-helices, and a cap domain of 5 α-helices. The catalytic triad is formed of a nucleophile (Asp), which attacks the epoxide and forms a covalent ester-intermediate, a catalytic His and a carboxylic acid which subsequently activate a water molecule and hydrolyze the ester bond to release the product [3,4,7-9]. The carboxylic acid occurs mostly after strand β7 (β7-position) of the α/β hydrolase fold, while in some bacterial EHs this residue is located after strand β6 (β6-position). The oxyanion hole consists of 2 residues, which donate their backbone amide protons to stabilize the negative charge of the transition state. EHs contain a HG**X**-motif [4], in which the **X** is the first oxyanion hole residue, and thus belong to the G**X**-type of hydrolases, which have first been classified for lipases and esterases [10]. The second oxyanion hole residue is a direct neighbor of the nucleophile. At least one Tyr is located in the cap domain and essential for activity. It is acting as proton donor and involved in substrate binding [11].

Mammalian EHs involved in detoxification are biochemically well characterized [12,13]. Recently, the interest in bacterial EHs is increasing, as they have been shown to catalyze the enantioselective conversion of industrially important epoxides and thus have a high potential as versatile biocatalysts for the preparation of optically pure epoxides and diols by kinetic resolution [14,15].

More than 100 EH genes have been sequenced and 3 EH structures have been determined experimentally: EHs from *Agrobacterium radiobacter* AD1[16,17], *Aspergillus niger* [18] and *Mus musculus* [19]. Up to now EHs have been classified in terms of activity and origin [13], but a systematic analysis of all EHs has not yet been performed. A systematic comparison of sequence, structure, and biochemical properties of all EHs should provide a basis to analyze the modular architecture of EHs, to classify known and newly sequenced EHs, to predict substrate specificity from their sequence, and to improve their biochemical properties by protein engineering.

Material and methods

Obtaining protein sequences

Sequences of EHs were retrieved from the NCBI-GenBank database [20]. As a first step, the database was searched for the keyword "epoxide". Sequence fragments and identical sequences, as determined by multisequence alignment with ClustalX 1.81 [21] were excluded. Subsequently, a phylogenetic tree of 71 remaining sequences was constructed with TREE-PUZZLE 5.0 using maximum-likelihood and quartet-puzzling [22]. 13 EH sequences (Table I) were used as templates for the construction of an EH database as described previously for the Lipase Engineering Database [23]: BLAST searches in GenBank were performed using the 13 representative EH sequences and a cutoff E-value of 10^{-10}. 75 sequences annotated as "EH", "putative EH" or "hydrolase" were chosen. Comparison of the 71 EH sequences of the keyword search and the 75 sequences of the database led to 95 different EH sequences.

(Table I)

Classification

A multisequence alignment of 95 EHs was performed with ClustalX 1.8.1 using default parameters (95 EH alignment) and is available as supplementary material. For definition of homologous families and superfamilies a phylogenetic tree was constructed using TREE-PUZZLE parallel version [22]. It was implemented on a Linux PC-cluster using Dual AMD Athlon MP 1800+ with Myrinet interconnect, with 16 processors per job. All phylogenetic trees were visualized using PHYLODENDRON and edited manually.

All sequences with a high similarity were assigned to a single homologous family. Homologous families with low but significant sequence similarities were grouped into a single superfamily, containing conserved positions of nucleophile, and catalytic His.

Remodeling of structures

The experimental structure of *Agrobacterium radiobacter* AD1 EH was remodeled using Swiss-PDB Viewer 3.7 (b2) [24] and the Swiss-Model [24] server, starting from the X-ray structure of the EH from *Agrobacterium radiobacter* AD1 (PDB entry: 1EHY) [16,17] and its sequence (EMBL database entry: CAA73331) [3] using the bromoperoxidase A2 from *Streptomyces aureofaciens* (PDB entry: 1BRO) [25] as template. In the structure of the *Agrobacterium radiobacter* EH, Met1 and the loop of residues 138-148 are missing. Probably due to crystal packing, the catalytic acid residue Asp246 is not in hydrogen bonding distance to the Nδ1 atom of the catalytic His, but is moved away into the solvent region. Instead Gln134 moved in and blocks the active site [16]. Therefore in addition residues 132-137 and 245-250 were remodeled using the bromoperoxidase A2 as template.

Structure comparison

Root mean square deviation (rmsd) between structures were obtained using Swiss-PDB Viewer 3.7. Only backbone atoms were used for the structural superimposition of EH structures. The percentage of residues involved in calculation of rmsd values is based on the longer sequence.

For the structurally known EHs from *Agrobacterium radiobacter*, *Aspergillus niger* and *Mus musculus*, the lengths of NC- and cap-loop were determined by structure comparison and assigning secondary structure elements by DSSP [26]. All NC-loops start directly after strand β6 and end before the first cap-helix. The cap-loop is located between the end of cap-helix 3 and start of cap-helix 4. As the positions of helices and strands are conserved in the 95 EH alignment, respective residues of start and stop of NC- and cap-loops were deduced from the alignment.

Homology modeling

The sequences of structurally unknown EHs were aligned to the most similar template sequence using Swiss-PDB Viewer 3.7 and adjusted manually to the ClustalX alignment of 95 EH sequences. Calculation of the homology models was done by the Swiss-Model server. The quality of the homology models was checked using the program "WhatCheck" [27] provided by Swiss-Model and by analyzing the position and orientation of the catalytic triad, the oxyanion hole and 2 conserved Tyr located in the cap domain. The structure models are accessible at our *Lipase Engineering Database*[23], http://www.led.uni-stuttgart.de .

Results

Sequence comparison

95 EHs were retrieved from GenBank. 59% of these EHs are putative proteins, mostly from bacterial, plant, and insect origin. Fungal and mammalian EHs are well investigated [13] and contain only 10% putative sequences. A multisequence alignment of 95 EHs demonstrates that the nucleophile and the catalytic His are conserved for all EHs, but not the catalytic acid residue at the β7-position. EHs were separated into 2 superfamilies, cytosolic and microsomal EHs. EHs of each superfamily were divided into separate homologous families, analyzed by multisequence alignment and phylogenetic analysis: 1) The microsomal EH superfamily (Fig. 1) contains 26 sequences grouped into 4 homologous families. These 4 families form 2 major branches. The higher organism branch contains a very closely related homologous family of Mammalian EHs and a more diverse homologous family of Insect EHs and 2 EHs from *C. elegans*. The microorganism branch contains a Fungal EH and a Bacterial EH homologous family, as well as 2 isolated bacterial sequences. Of one member, the fungal EH from *Aspergillus niger*, the structure is known. As all microsomal EHs of higher organisms are membrane bound, Mammalian and Insect microsomal EHs contain an N-terminal membrane anchor (Fig. 3b). In contrast, microbial EHs of this superfamily are soluble proteins and therefore lack the anchor [8]. 2) The cytosolic EH superfamily (Fig. 2) contains 69 sequences grouped into 8 homologous families, which form 2 major branches. The higher organism branch contains a very closely related homologous family of Mammalian EHs, 2 more diverse homologous families of Plant EHs, and a Bacterial family related to EHs of higher organisms. The microrganism branch contains 4 homologous families, a

diverse set of bacterial sequences and an isolated EH from *C. elegans*. The majority of bacterial sequences are found in this superfamily. Of 2 members (the Mammalian EH from *Mus musculus* and the Bacterial EH from *Agrobacterium radiobacter*) the structure is known. All cytosolic EHs are soluble enzymes. The EHs of 2 bacterial homologous families contain their catalytic acid residues (Asp) at the β6-position (β6 EHs) or at both positions (β6/β7 EHs), including the well characterized EH from *Corynebacterium sp.,* for which Asp123, located after strand β6, was predicted as the catalytic acid residue [4].

(Figures 1 and 2)

EH structures

The nucleophile (Asp) and catalytic His are located in the predominantly hydrophobic region between core and cap domain. The nucleophile is situated at a sharp nucleophilic elbow after the central strand β5, the catalytic His is located after strand β8. The catalytic acid residue (Asp or Glu) is located after strand β7 or in a few cytosolic bacterial EHs after strand β6 of the α/β-hydrolase fold [18]. The catalytic triad of cytosolic EHs is similar to microsomal EHs [17]. At a first glance, the 3 published EH structures (*Agrobacterium radiobacter* AD1, PDB entry: 1EHY; *Mus musculus* PDB entry: 1CQZ; and *Aspergillus niger* PDB entry: 1QO7) seem to be rather different in size and in shape (Fig. 3 a). However, a detailed structure comparison supported by multisequence alignment of 95 sequences identified 3 conserved and 3 variable regions (Fig. 3 b): 1) The N-terminal region is highly variable and its structure differs for the superfamilies. Plant EHs lack this N-terminal domain, whereas in cytosolic Mammalian EHs, the structure of this cytosolic domain is similar to the structure of

haloacid dehalogenases with a recently detected phosphatase activity [28,29] and connected to the core by a linker [19]. Instead, all microsomal EHs contain a microsomal domain. Mammalian and insect microsomal EHs have an additional N-terminal membrane anchor. Most bacterial EHs lack this N-terminal region. 2) The N-terminal region is followed by the N-terminal half of the conserved core domain. It constitutes the first half of the central α/β hydrolase domain, and consists of 6 β-strands and 4 α-helices. This catalytic domain also contains the nucleophile and the β6-position. 3) A variable NC-loop links the conserved N-terminal catalytic domain and the conserved **c**ap domain. The NC-loop is predominantly hydrophobic; its length ranges from 16 to 57 residues. In the EHs from *Mus musculus* (23 residues) and *Aspergillus niger* (35 residues), the NC-loop forms an α-helix, but not in the *Agrobacterium radiobacter* EH (19 residues). 4) For all EHs the conserved cap domain consists of 5 helices in an uteroglobin-like fold [30]. It consists of 2 layers: The upper layer is formed by helices 1, 2, 3 with a trapeze-like structure. The lower layer (helices 4 and 5) has a V-shape with an angle of about 100°. In the 3 structures, helices 1, 2, 3, 4 have similar length (helix 1: 5-8, helix 2: 7-10, helix 3: 14 and helix 5: 8-12 residues). Helix 5 has 11 residues (*Agrobacterium radiobacter* and *Aspergillus niger*, respectively), but only 4 residues in the EH from *Mus musculus*. The cap domain contains a conserved Tyr in helix 5 and in most EHs a second Tyr in helix 1. 5) A variable cap-loop is inserted into the cap domain between helix 3 and 4. The length of this loop varies: 8 residues for *Aspergillus niger* EH, 9 residues for *Agrobacterium radiobacter* AD1 EH, and 36 residues for *Mus musculus* EH. For the 3 known structures the hydrophobicity increases with length: the *Aspergillus niger* EH contains a hydrophilic cap-loop (25% hydrophobic residues), for the *Agrobacterium radiobacter* EH the loop is rather hydrophobic (43% hydrophobic residues), and the

loop of *Mus musculus* EH is mostly hydrophobic (63% hydrophobic residues). 6) The cap domain is followed by the C-terminal half of the conserved core domain. It constitutes the second half of the central α/β hydrolase domain and consists of 2 β-strands and 2 α-helices. It contains the catalytic His and the β7-position, responsible for the hydrolysis of the ester intermediate.

(Figure 3)

Sequence similarities between the structurally known EHs are moderate, while the structure of the core and the cap domain is highly conserved. However, sequence and structure similarity are not strictly coupled: For the pair mouse EH and *Agrobacterium* EH, the sequence similarity between the core domains is 26%, the corresponding root mean square deviation (rmsd) of the backbone atoms is 1.3 Å (including 90% of all residues). In contrast, *Aspergillus* EH and *Agrobacterium* EH show sequence similarity of only 17% identity and a rmsd of also 1.3 Å (including 72% of all residues). Apparently the structure of the α/β hydrolase fold is very similar for all EHs. However, other members of the α/β hydrolase fold family may deviate in structure and even in the number of strands (*Bacillus subtilis* lipase A [31]) or their topology (*Rhizomucor miehei* lipase [32]).

95 EH alignment

The modular architecture of the 3 EHs of known structure is consistent with a multisequence alignment of 95 EHs. Although the alignment includes EHs from 2 superfamilies, and the position of the catalytic acid residue varies, the boundaries between the structural modules and the sequences of core and cap domain are

conserved. 5 residues are absolutely conserved for all EHs: 1) 2 members of the catalytic triad, the nucleophile and the catalytic His, which are in direct contact with the substrate. The third member of the catalytic triad, the catalytic acid residue, shows plasticity: in most EHs it occurs after the central strand β7, in some EHs however after β6. Interestingly a few bacterial EHs contain acid residues at both positions. The side chains of both acid residues point towards the catalytic His and are in hydrogen bond distance to stabilize the positive charge of the imidazole-ring. 2) The oxyanion hole is formed by the backbone amide hydrogen atoms of 2 residues. The first is part of the HG**X**-Motif, the second is following the nucleophile. 3) All EHs contain a conserved Tyr in helix 5 of the cap domain. Most EHs contain a second Tyr residue in helix 1 of the cap domain. This Tyr does not align in the multisequence alignment of 95 EHs and is even lacking in individual EHs of both superfamilies and in all β6 EHs.

Clustering by loop length

Both superfamilies contain bacterial EHs, but only of one bacterial EH the structure has been solved (*Agrobacterium radiobacter* EH, a member of the cytosolic EH superfamily). As their sequences are highly diverse and hardly cluster, an additional criterion is needed for classification. Since length and conformation of the NC- and the cap-loop are the major differences between EHs, loop lengths and properties were analyzed for classification and structure prediction. To systematically analyze loop lengths for all EHs, their first and last residues were derived for all 95 EH sequences by sequence alignment. Based on loop length, the members of the microsomal and the cytosolic EH superfamilies form 3 clusters (Fig. 4). Individual clusters are continuously populated and clearly separated from each other: 1) EHs of cluster I have medium sized NC- and long cap-loops (16-40 and 31-59 residues, respectively)

and include 3 homologous families of the cytosolic EHs of plant and mammalian origin and a bacterial family related to EHs of higher organisms. Cluster I contains the EH from *Mus musculus* with known X-ray structure, which has NC- and cap-loops of 23 and 35 residues, respectively. Mammalian EHs have loops of similar length (NC-loop: 23, cap-loop: 35 or 36 residues), Plant EHs show more variability (NC-loop: 16-25, cap-loop: 31-35 residues) and the bacterial EHs of cluster I show variability in both loops (NC-loop: 18-40, cap-loop: 38-59 residues). 2) EHs of cluster II are characterized by short NC- and cap-loops (18-25 and 5-12 residues, respectively). It includes the bacterial EHs of the cytosolic EH superfamily, excluding the mammalian related bacterial EHs. Of one member, the EH from *Agrobacterium radiobacter*, the structure is known (NC-loop: 19 and cap-loop: 9 residues). Cluster II is clearly separated from cluster I, although their members are EHs of the same superfamily. Cluster II is clearly separated from cluster I, as can be seen in the corresponding phylogenetic tree (Fig. 2), where also the majority of the bacterial EHs is clearly separated from all other cytosolic EHs. Although the cluster II bacterial EHs are diverse in sequence, they are closely related in terms of loop length. 3) All microsomal EHs, including the bacterial EHs, form a single cluster III. The length of the NC-loops varies widely (21-57 residues), while the cap-loops of all 4 homologous families are short to medium sized (8-19 residues, respectively). Cluster III contains the X-ray structure of *Aspergillus niger* EH (NC-loop: 35, cap-loop: 8 residues).

(Figure 4)

Secondary structure of NC-loops

The 3 clusters have overlapping NC-loop lengths. The NC-loop of the *Agrobacterium radiobacter* EH of 19 residues contains no helical elements, whereas the EH of *Mus musculus* and *Aspergillus niger* contain NC-loops (23 and 36 residues, respectively), which form a helix. For all 3 NC-loops, the length of the non-helical part of the loop is 16–19 residues; all residues above this threshold are packed into a helix. Therefore, the formation of an α-helix in the NC-loop is expected to occur in all clusters and only depend on loop length (Fig. 4): 1) For short NC-loops up to 19 residues, no helix is expected to form (e.g. 19 residues for *Agrobacterium radiobacter* EH); 2) for medium sized NC-loops (20 to 32 residues), the helix is expected to be short (e.g. 23 residues for *Mus musculus* EH); 3) for long NC-loops (33 to 40 residues), the helix is expected to be long (e.g. 35 residues for *Aspergillus niger* EH); 4) for very long NC-loops (longer than 40 residues), no prediction can be given due to lack of experimental structure data. Short and medium NC-loops occur in cluster I and II, long and very long loops in cluster III. The variable NC-loop is flanked by 2 structurally conserved regions, the N-terminal catalytic domain and the cap domain. Thus, the NC-loop seems to fold independently from the conserved domains of the protein, the α-helical content seems to exclusively depend on loop length.

Homology modeling

Homology modeling is based on a reliable alignment of the sequences of target and template. Using pairwise alignment between the two sequences, a sequence identity of more than 30% is prerequisite. Due to the high diversity of EHs in sequence space and the existence of only 3 template structures, the number of reliable pairwise alignments is limited to only 14% of all 95 EHs. In order to extent homology modeling to all

EHs, the 95 EH alignment was used as reference for homology modeling. The 3 experimental EH structures demonstrate that structure is highly conserved between the superfamilies with the exception of the NC- and cap-loop which may vary considerably. While the conserved α/β hydrolase fold and the cap region superpose well in the sequence alignment and therefore can be reliably modeled, the feasibility of modeling the loops depends on their respective lengths and was investigated for each loop cluster. For representative EHs of each cluster homology models were derived and deposited at http://www.led.uni-stuttgart.de.

Cluster I: For most EHs of cluster I, the *Mus musculus* EH can be used as template, but for some EHs the loops could not be modeled based on this template: two Plant EHs contain NC-loops of 16 and 19 residues and therefore are predicted to be non-helical. While the NC-loop of *Mus musculus* EH (23 residues) is helical, the respective loop of *Agrobacterium radiobacter* EH (19 residues) is of appropriate length and therefore can be used as template instead. In addition, several bacterial EHs of this cluster contain NC-loops longer than 28 residues, for which the NC-loop of *Aspergillus niger* EH (35 residues) can be used as template. However, some bacterial EHs have cap-loops longer than 40 residues, for which no template is available. In summary, the structure of 94% of all cluster I EHs can be completely predicted. As representative models, the structure of EHs from *Glycine max* (19% overall identical residues, 30% identical residues *Mus musculus* EH, excluding the N-terminal domain) and from *Streptomyces coelicolor* (20% identical residues) are available.

Cluster II: Since all EHs of cluster II contain similar loop lengths, all structures of the cluster II EHs can be predicted. For one EH with a NC-loop longer than 20 residues, the NC-loop of the *Mus musculus* EH can be used as template. Most EHs of this

cluster have their catalytic acid residue in β7-position, but some at β6-position. For both types of EHs, *Agrobacterium radiobacter* EH can be used as a template: it has an appropriate NC- and cap-loop length (19 and 9 residues, respectively) and luckily, this EH has an Asp in both positions, which are in hydrogen bond distance to the catalytic His. As representative models, the structure of EHs from *Mycobacterium tuberculosis* (26% identical residues, Asp in β7-position) and from *Corynebacterium sp.*(25 % identical residues, Asp in β6-position) are available.

Cluster III: For 50 % of all microsomal EHs, the cap-loop can be modeled based on *Aspergillus niger* EH as template, while for only 19% of all microsomal EHs the structure of both loops can be predicted. For Mammalian EHs, both loops are too long to be modeled, for Insect EHs and some Fungal EHs the NC-loop is much longer than in the template. As representative model, the structure of the EH from *Mesorhizobium loti* (23% identical residues) is available.

(Table II)

Discussion

The role of structural modules

The α/β hydrolase fold is among the most frequent folds and includes EHs and other hydrolases like lipases, acetylcholinesterases, carboxypeptidases, and haloperoxidases [30]. Even though these enzymes catalyze different types of reactions and accept different substrates, all α/β hydrolases contain a similar catalytic triad: nucleophile-His-acid [30].

In the highly conserved framework of architecture and geometry of the catalytic machinery, a broad range of variations are observed. EHs show 2 different types of

such variations in the active site: variation in the position of the catalytic acid residue, which can be located either after the central β-strand β6 or after β7 [4], and the existence of a second Tyr in the cap domain.

The cap domains of EHs and the homologous haloalkane dehalogenases contain residues that are catalytically relevant and involved in substrate binding [11] and, for haloalkane dehalogenases, in halide binding [33]. For haloalkane dehalogenases it is proposed that the cap domain is directly involved in the reaction and relevant for substrate specificity [30].

Cytosolic and microsomal EHs differ in their substrate spectra [13]. A narrow substrate channel is visible between NC- and cap-loop. This is supported by the high mobility of the NC-loop in *Agrobacterium radiobacter* EH as concluded from high B-factors in the crystal structure [16], and the high flexibility of the NC-loop of haloalkane dehalogenases in molecular dynamics simulations [34]. Because of its flexibility the NC-loop could move aside and open the way inside the molecule. In addition, cap-helix α3 , which is located above the NC-loop and next to the cap-loop, also was supposed to be mobile, as described for the structurally similar haloalkane dehalogenases [35]. Both NC- and cap-loop differ clearly in length for the 3 known structures. Therefore, differences in length of NC- and cap-loop may have a direct effect on shape and accessibility of the active site and could mediate the different substrate spectra of EHs: while microsomal EHs containing long NC-loops are able to convert space filling polycyclic aromatic hydrocarbons [13], cytosolic Mammalian and Plant EHs convert epoxy fatty acids. They contain a long cap-loop, which is located above the nucleophile an forms a hydrophobic tunnel. The binding pocket of these EHs has a L-shape [19] similar to lipases which convert long chain fatty acids [36].

Of 78% of all 95 EHs the complete structure can be predicted using the insights resulting from the systematic analysis. In contrast, only 14% of these EHs can be predicted using pairwise alignments. The number of structures possible to predict varies for the 3 clusters: Whereas 94% of all cluster I and 100% of all cluster II EHs are predictable, only 19% of the microsomal EHs of cluster III can be completely modeled. Due to their medical importance, human microsomal EHs are well investigated enzymes [13], which contain very long NC-loops. A crystal structure of such an enzyme is urgently needed to cover the white spots in the structure map. In summary, the structure of all α/β hydrolase domains, 78% of NC- and 89% of cap-loops of all EH sequences can be predicted by homology modeling.

Functional families

The separation into superfamilies based on sequence similarity and clustering by loop length correlates with substrate specificity and metabolic function of EHs (Table III): 1) All **cytosolic EHs of higher organisms** have long cap-loops and medium-sized NC-loops. They are active towards a common substrate class, aliphatic epoxides, and are involved in the metabolism of fatty acids [13]. The homologous family of Mammalian EHs are involved in the xenobiotic metabolism and the degradation of endogenously derived epoxy fatty acids, while the family of Plant EHs play a central role in the biosynthesis of essential aliphatic cuticular compounds, like epoxy stearic acids [37] and in detoxification of epoxy fatty acids in plant seeds [7]. 2) All **bacterial cytosolic EHs of cluster II** have short cap- and NC-loops. Several members of this family have been shown to accept small substrates like styrene oxide, mono- and disubstituted epoxides [15].

3) **Microsomal EHs** have short to medium cap-loops and medium to long NC-loops. They are mainly involved in the metabolism of polycyclic aromatic compounds. The family of Insect EHs are involved in regulation of juvenile hormones [38] by degradation of polycyclic epoxy sesquiterpenes, while the well investigated mammalian microsomal EHs are involved in the bioactivation of carcinogenic polycyclic hydrocarbons and the detoxification of epoxide intermediates [13]. While most of the microsomal EHs have short cap-loops, the mammalian microsomal EHs have medium-sized cap loops. Interestingly, this family is also able to convert epoxy fatty acids [39], like cytosolic EHs of higher organisms. From these data it seems that long cap-loops lead to the ability to convert aliphatic substrates, long NC-loops mediate conversion of polycyclic aromatic hydrocarbons, while EHs with short cap- and NC-loops prefer small substrates. According to this rule, the substrate specificity of non-characterized bacterial EHs of cluster I and III can be predicted. Cluster I EHs contain long cap-loops and therefore are expected to convert epoxy fatty acids. Cluster III EHs contain long cap-loops and therefore are expected to convert epoxy fatty acids. This observed relationship between loop length and substrate specificity is further supported by experimental data on mutants of a haloalkane dehalogenase [40]. For this homologous enzyme, varying the length of the NC-loop led to variations in chain length specificity towards various chloroalkanes.

(Table III)

Although activity towards epoxy fatty acids and polycyclic aromatic hydrocarbons is essential for an organism, only mammalian genomes contain EHs of both cytosolic and mammalian EH superfamilies. In contrast, yeasts, fungi, insects, and plants

encode only a single EH superfamily. As genomes of plants contain a wide range of epoxy steroids and diterpenoids [41], microsomal EHs are lacking. However, glutathione S-transferase (GST) was found in many plants like *Arabidopsis thaliana* or *Zea mays* [42]. It is able to open oxirane rings of epoxides [43] and thus might be a plant specific pathway of degrading epoxy polycyclic aromatic hydrocarbons. Similarly, yeasts and fungi exclusively encode microsomal EHs. The activity of the lacking cytosolic EHs is probably taken over by other enzymes: *Saccharomyces cerevisiae* is able to convert the epoxy fatty acid leukotriene using a leukotriene A(4) hydrolase [44], which has no sequence similarity to EHs and belongs to the family of zinc metalloproteases with both activities, an epoxide hydrolase and a protease activity [45].

Interestingly, genomes of some bacterial genera (*Mycobacterium*) or species (*Caulobacter crescentus, Mesorhizobium loti, Sinorhizobium meliloti*) contain EHs of both superfamilies. This broad distribution of bacterial EH sequences reflects the broad substrate spectrum to which bacteria are exposed [46]. This seems also to apply to other enzyme families like cellulases. In general, bacteria show a high diversity caused by the ecological habitats occupied by these organisms [47].

Although EHs are highly diverse in sequence, their structure is highly conserved. Therefore structure prediction is possible despite the fact that sequence similarities generally is below the threshold for reliable homology modeling. In addition, substrate specificity seems to be dominated by the length of two variable loops. This systematic analysis demonstrates the modular architecture of EHs, which opens the way to a deeper understanding of structure and function of EHs and other homologous α/β hydrolases.

Supplementary material

The annotated 95 EH alignment and homology models of 5 representative EHs are available as supplementary material.

Acknowledgement

We thank the BASF AG, Ludwigshafen, Germany, for financial support.

References

1. Armstrong RN. Enzyme-catalyzed detoxication reactions: mechanisms and stereochemistry. CRC Crit Rev Biochem 1987;22:39-88.
2. Beetham JK, Grant D, Arand M, Garbarino J, Kiyosue T, Pinot F, Oesch F, Belknap WR, Shinozaki K, Hammock BD. Gene evolution of epoxide hydrolases and recommended nomenclature. DNA Cell Biol 1995;14:61-71.
3. Rink R, Fennema M, Smids M, Dehmel U, Janssen DB. Primary structure and catalytic mechanism of the epoxide hydrolase from *Agrobacterium radiobacter* AD1. J Biol Chem 1997;272:14650-14657.
4. Misawa E, Chan Kwo Chinon CK, Archer IV, Woodland MP, Zhou NY, Carter SF, Widdowson DA, Leak DJ. Characterisation of a catabolic epoxide hydrolase from a *Corynebacterium* sp. Eur J Biochem 1998;253:173-183.
5. Debernard S, Morisseau C, Severson TF, Feng L, Wojtasek H, Prestwich GD, Hammock BD. Expression and characterization of the recombinant juvenile hormone epoxide hydrolase (JHEH) from *Manduca sexta*. Insect Biochem Mol Biol 1998;28:409-419.
6. Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J et al. The alpha/beta hydrolase fold. Protein Eng 1992;5:197-211.
7. Arahira M, Nong VH, Udaka K, Fukazawa C. Purification, molecular cloning and ethylene-inducible expression of a soluble-type epoxide hydrolase from soybean (*Glycine max* [L.] Merr.). Eur J Biochem 2000;267:2649-2657.
8. Arand M, Hemmer H, Durk H, Baratti J, Archelas A, Furstoss R, Oesch F. Cloning and molecular characterization of a soluble epoxide hydrolase from *Aspergillus niger* that is related to mammalian microsomal epoxide hydrolase. Biochem J 1999;344:273-280.
9. Arand M, Wagner H, Oesch F. Asp333, Asp496, and His523 form the catalytic triad of rat soluble epoxide hydrolase. J Biol Chem 1996;271:4223-4229.
10. Pleiss J, Fischer M, Peiker M, Thiele C, Schmid RD. Lipase engineering database - Understanding and exploiting sequence-structure-function relationships. J Mol Cat Enzym B 2000;10:491-508.
11. Argiriadi MA, Morisseau C, Goodrow MH, Dowdy DL, Hammock BD, Christianson DW. Binding of alkylurea inhibitors to epoxide hydrolase implicates active site tyrosines in substrate activation. J Biol Chem 2000;275:15265-15270.
12. Oesch F. Mammalian epoxide hydrase. Xenobiotica 1972;3:305-340.
13. Fretland AJ, Omiecinski CJ. Epoxide hydrolases: biochemistry and molecular biology. Chem Biol Interact 2000;129:41-59.
14. Swaving J, de Bont J. Microbial transformation of epoxides. Enzyme Microb Technol 1998;22:19-26.
15. Steinreiber A, Faber K. Microbial epoxide hydrolases for preparative biotransformations. Curr Opin Biotechnol 2001;12:552-558.
16. Nardini M, Ridder IS, Rozeboom HJ, Kalk KH, Rink R, Janssen DB, Dijkstra BW. The x-ray structure of epoxide hydrolase from *Agrobacterium radiobacter* AD1. J Biol Chem 1999;274:14579-14586.

17. Nardini M, Rink R, Janssen DB, Dijkstra BW. Structure and mechanism of the epoxide hydrolase from Agrobacterium radiobacter AD1. J. Mol. Cat. Enzym. B 2001;11:1083-1090.
18. Zou J, Hallberg BM, Bergfors T, Oesch F, Arand M, Mowbray SL, Jones TA. Structure of *Aspergillus niger* epoxide hydrolase at 1.8 A resolution: implications for the structure and function of the mammalian microsomal class of epoxide hydrolases. Structure 2000;8:111-122.
19. Argiriadi MA, Morisseau C, Hammock BD, Christianson DW. Detoxification of environmental mutagens and carcinogens: structure, mechanism, and evolution of liver epoxide hydrolase. Proc Natl Acad Sci U S A 1999;96:10637-10642.
20. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res 2002;30:17-20.
21. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The Clustal_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;24:4876-4882.
22. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 2002;18:502-504.
23. Fischer M, Pleiss J. The Lipase Engineering Database: a navigation and analysis tool for protein families. Nucleic Acids Res 2003;31:319-321.
24. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714-2723.
25. Pfeifer O, Pelletier I, Altenbuchner J, van Pee KH. Molecular cloning and sequencing of a non-haem bromoperoxidase gene from *Streptomyces aureofaciens* ATCC 10762. J Gen Microbiol 1992;138:1123-1131.
26. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577-2637.
27. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature 1996;381:272.
28. Cronin A, Mowbray S, Durk H, Homburg S, Fleming I, Fisslthaler B, Oesch F, Arand M. The N-terminal domain of mammalian soluble epoxide hydrolase is a phosphatase. Proc Natl Acad Sci U S A 2003;100:1552-1557.
29. Newman JW, Morisseau C, Harris TR, Hammock BD. The soluble epoxide hydrolase encoded by EPXH2 is a bifunctional enzyme with novel lipid phosphate phosphatase activity. Proc Natl Acad Sci U S A 2003;100:1558-1563.
30. Holmquist M. Alpha/Beta-hydrolase fold enzymes: structures, functions and mechanisms. Curr Protein Pept Sci 2000;1:209-235.
31. van Pouderoyen G, Eggert T, Jaeger KE, Dijkstra BW. The crystal structure of *Bacillus subtilis* lipase: a minimal alpha/beta hydrolase fold enzyme. J Mol Biol 2001;309:215-226.
32. Derewenda ZS, Derewenda U, Dodson GG. The crystal and molecular structure of the *Rhizomucor miehei* triacylglyceride lipase at 1.9 A resolution. J Mol Biol 1992;227:818-839.

33. Damborsky J, Koca J. Analysis of the reaction mechanism and substrate specificity of haloalkane dehalogenases by sequential and structural comparisons. Protein Eng 1999;12:989-998.
34. Otyepka M, Damborsky J. Functionally relevant motions of haloalkane dehalogenases occur in the specificity-modulating cap domains. Protein Sci 2002;11:1206-1217.
35. Pikkemaat MG, Linssen AB, Berendsen HJ, Janssen DB. Molecular dynamics simulations as a tool for improving protein stability. Protein Eng 2002;15:185-192.
36. Pleiss J, Fischer M, Schmid RD. Anatomy of lipase binding sites: the scissile fatty acid binding site. Chem Phys Lipids 1998;93:67-80.
37. Blée E, Schuber F. Biosynthesis of cutin monomers - involvement of a lipoxygenase/peroxygenase pathway. Plant J 1993;4:113-123.
38. Wojtasek H, Prestwich GD. An insect juvenile hormone-specific epoxide hydrolase is related to vertebrate microsomal epoxide hydrolase. Biochem Biophys Res Commun 1996;220:323-329.
39. Summerer S, Hanano A, Utsumi S, Arand M, Schuber F, Blée E. Stereochemical features of the hydrolysis of 9,10-epoxystearic acid catalysed by plant and mammalian epoxide hydrolases. Biochem J 2002;366:471-480.
40. Pries F, van den Wijngaard AJ, Bos R, Pentenga M, Janssen DB. The role of spontaneous cap domain mutations in haloalkane dehalogenase specificity and evolution. J Biol Chem 1994;269:17490-17494.
41. Abdel-Mogib M, Albar HA, Batterjee SM. Chemistry of the genus *Plectranthus*. Molecules 2002;7:271-301.
42. Labrou NE, Mello LV, Clonis YD. Functional and structural roles of the glutathione-binding residues in maize (*Zea mays*) glutathione S-transferase I. Biochem J 2001;358:101-110.
43. Eaton DL, Bammler TK. Concise review of the glutathione S-transferase and their significance to toxicology. Toxicol Sci 1999;49:156-164.
44. Kull F, Ohlson E, Lind B, Haeggstrom JZ. *Saccharomyces cerevisiae* leukotriene A4 hydrolase: formation of leukotriene B4 and identification of catalytic residues. Biochemistry 2001;40:12695-12703.
45. Haeggstrom JZ. Structure, function, and regulation of leukotriene A4 hydrolase. Am J Respir Crit Care Med 2000;161:25-31.
46. Mechichi T, Stackebrandt E, Gad'on N, Fuchs G. Phylogenetic and metabolic diversity of bacteria degrading aromatic compounds under denitrifying conditions, and description of *Thauera phenylacetica* sp. nov., *Thauera aminoaromatica* sp. nov., and *Azoarcus buckelii* sp. nov. Arch Microbiol 2002;178:26-35.
47. Bjedov I, Tenaillon O, Gerard B, Souza V, Denamur E, Radman M, Taddei F, Matic I. Stress-induced mutagenesis in bacteria. Science 2003;300:1404-1409.

Tables

Table I: Representative EH sequences used for BLAST search to construct the database.

Group	GenBank	Organism	Putative
Microsomal	AAB18243	*Trichoplusia ni*	
C. elegans mic.	NP_505811	*C. elegans*	
Mammal cyt	NP_001970	*Homo sapiens*	
Plant1, cyt	AAB02006	*Nicotiana tabacum*	
Plant2,cyt	CAA55294	*Glycine max*	
Bac1, cyt	1EHYA	*Agrobacterium radiobacter* AD1	
Bac2, cyt	NP_103292	*Mesorhizobium loti*	
Bac3, cyt	C83216	*Pseudomonas aeruginosa*	putative
Bac4, cyt	CAC37878	*Streptomyces coelicolor*	putative
Bac5, cyt	NP_334552	*Mycobacterium tuberculosos* CDC1551	
Bac6, cyt	NP_396231	*Agrobacterium tumefaciens* C58	
Bac7, cyt	CAA11900	*Corynebacterium sp.*	
Bac8, cyt	NP_107141	*Mesorhizobium loti*	

Table II: Target and template EHs for homology modeling

Organism	Sequence	Template	Sequence identity	Catalytic triad
Homo sapiens	AAG14967	*Mus musculus*	72%	D335, D496, H524
Glycine max	CAA55294	*Mus musculus*	19%	D126, D285, H320
Streptomyces coelicolor	T36559	*Mus musculus*	20%	D126, D300, H331
Mycobacterium tuberculosis	NP_334552	*Agrobacterium radiobacter*	26%	D108, (D246), H274
Mesorhizobium loti	NP_107140	*Aspergillus niger*	23%	D226, E393, (H420)

Table III: Classification of EHs by loop length and substrate specificity

EH superfamily	Organism	Cluster	NC-loop	Cap-loop	Substrate
cytosolic	plants/mammals	I	short-medium	long	epoxy fatty acids [c]
	bacteria	I	short-medium	long	epoxy fatty acids [p]
	bacteria	II	short	short	small substrates [c]
microsomal	insects	III	long–very long	short	polycyclic aromatic hydrocarbons [c]
	mammals	III	long–very long	medium	polycyclic aromatic hydrocarbons [c]
	yeasts/fungi	III	long–very long	short	polycyclic aromatic hydrocarbons [c]
	bacteria	III	long	short	polycyclic aromatic hydrocarbons [p]

[c]: biochemically characterized
[p]: putative (predicted by loop length)

Figure captions

Fig.1: Phylogenetic tree of microsomal EHs (cluster III). The higher organism branch (upper half) consists of the 2 homologous families (Insect EHs and Mammalian EHs) and 2 EHs from *C. elegans*. The microorganism branch (lower half) consists of 2 homologous families (Fungal EHs and Bacterial EHs) and 2 isolated bacterial EHs. The structurally known EH from *Aspergillus niger* is marked with a star.

Fig.2: Phylogenetic tree of cytosolic EHs. The higher organism branch (cluster I, upper half) consists of 4 homologous families (Mammalian EHs, Bacterial EHs related to higher organisms, 2 Plant EHs). The microorganism branch (cluster II, lower half) contains a diverse set of bacterial EHs including 4 homologous families. The structurally known EHs from *Mus musculus* and *Agrobacterium radiobacter* are marked with stars.

Fig.3:

a) The three solved EH structures from *Mus musculus* (upper left), *Aspergillus niger* (upper right), and *Agrobacterium radiobacter* AD1 (lower right): the core domain consists of the N- (blue) and C- (yellow) terminal catalytic domain, and the cap domain (red) which includes the cap-loop (violet). Core and cap domains are connected by the NC-loop (brown). The *Mus musculus* EH contains an additional cytosolic domain (dark green) connected by a linker (pink), the *Aspergillus niger* EH a microsomal domain (light green). Position and number of cap-helices are labeled for the *Agrobacterium radiobacter* EH as α1-α5.

b) Modular structure of homologous EH families (Plant, cytosolic Mammalian, Fungal, Insect, and microsomal Mammalian EHs): cytosolic domain (dark green) and linker (pink) of cytosolic EHs; microsomal domain (light green) of the microsomal EHs; membrane anchor (black) of Insect and Mammalian EHs; N-terminal catalytic domain (blue); NC-loop (brown) of variable length from 16 to57 residues; cap domain (red) with a variable cap-loop (violet) from 5 to 59 residues inserted; C-terminal catalytic domain (yellow). Bacterial EHs consist of many homologous families.

Fig.4: Loop lengths of NC- and cap-loop: Cluster I (cytosolic Mammalian and Plant EHs and Bacterial EHs related to EHs from higher organisms), cluster II (Bacterial EHs), and cluster III (microsomal EHs) according to the 2 superfamilies of cytosolic EHs (upper left) and microsomal EHs (lower right). EHs with experimentally determined structure are indicated by an arrow. The ruler below the graph indicates the regions of experimentally determined or proposed helix length of the NC-loop.

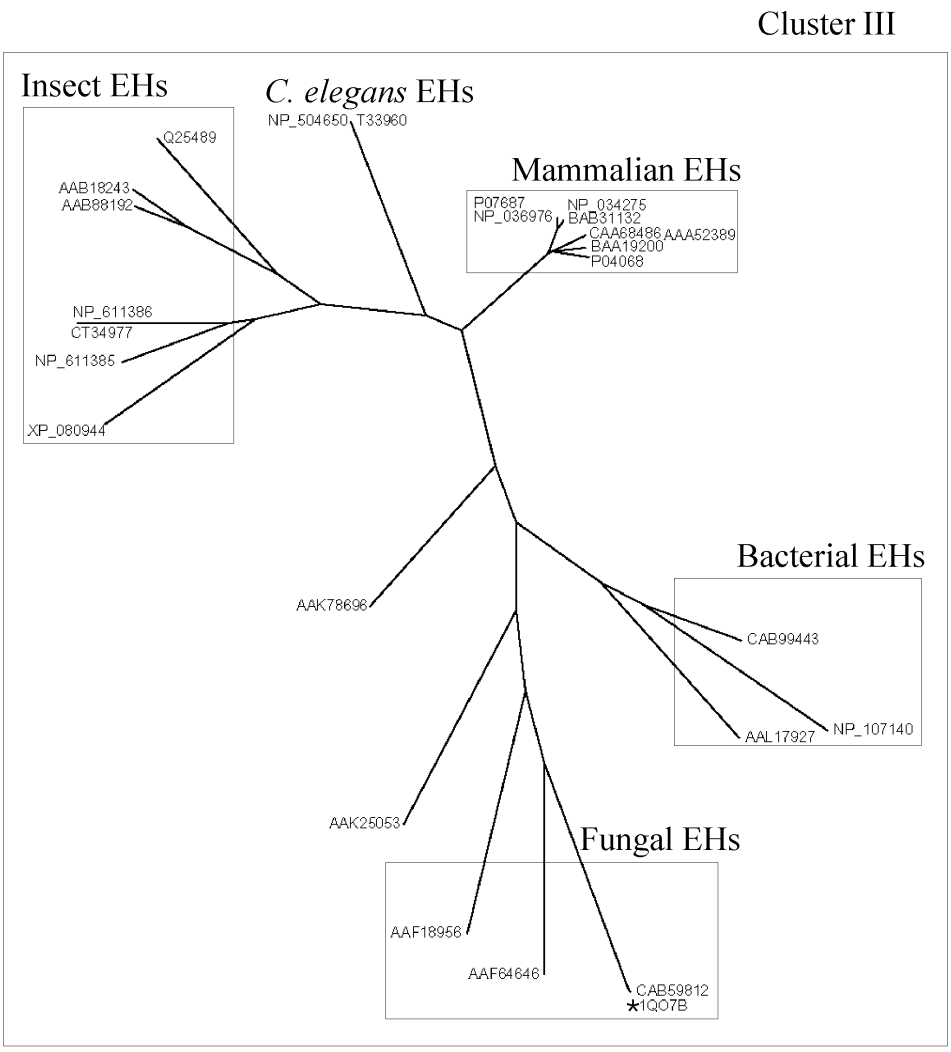

Cluster III
Insect EHs
C. elegans EHs
NP_504650
T33960
Mammalian EHs
Q25489
AAB18243
AAB88192
P07687
NP_036976
NP_034275
BAB31132
CAA68486
AAA52389
BAA19200
P04068
NP_611386
CT34977
NP_611385
XP_080944
Bacterial EHs
AAK78696
CAB99443
AAL17927
NP_107140
AAK25053
Fungal EHs
AAF18956
AAF64646
CAB59812
1QO7B
0.1

Figure 1

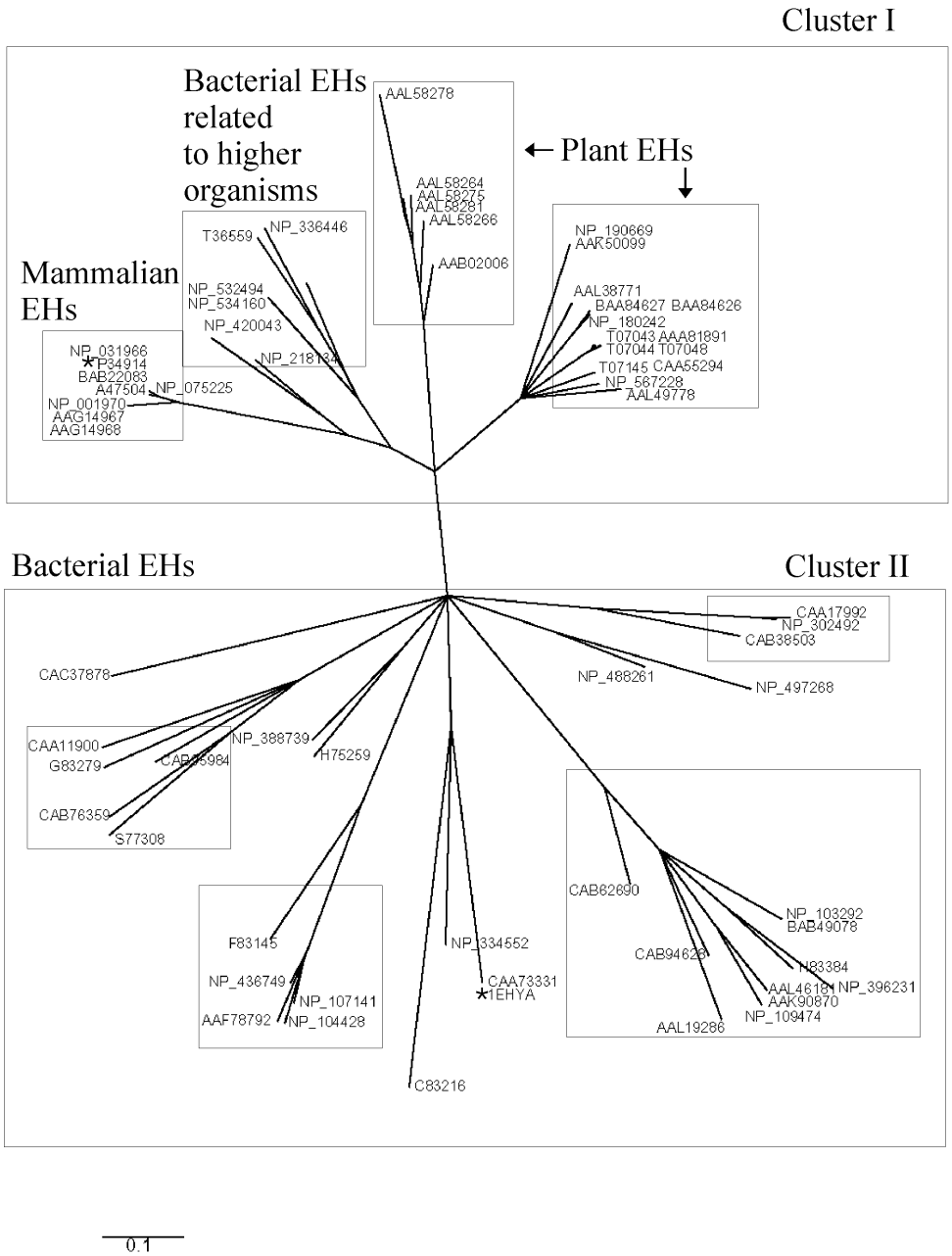
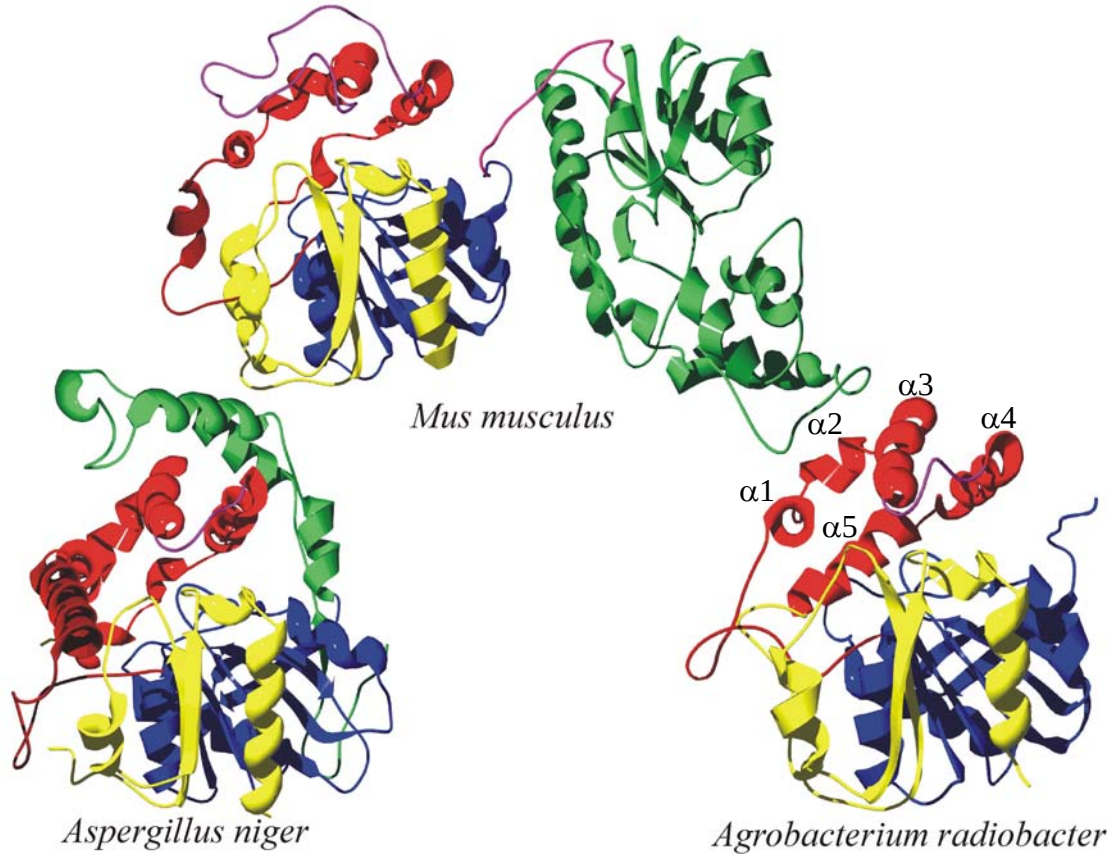

Cluster I
Bacterial EHs related to higher organisms
Plant EHs
Mammalian EHs
AAL58278
AAL58264
AAL58275
AAL58281
AAL58266
AAB02006
T36559
NP_338446
NP_532494
NP_534160
NP_420043
NP_218134
NP_031966
P34914
BAB22083
A47504
NP_075225
NP_001970
AAG14967
AAG14968
NP_190669
AAR50099
AAL38771
BAA84627 BAA84626
NP_180242
T07043 AAA81891
T07044 T07048
T07145 CAA55294
NP_567228
AAL49778
Bacterial EHs
Cluster II
CAA17992
NP_302492
CAB38503
CAC37878
NP_488261
NP_497268
CAA11900
G83279
CAB76359
S77308
NP_388739
H75259
F83145
NP_436749
NP_107141
AAF78792
NP_104428
NP_334552
CAA73331
1EHYA
C83216
CAB62690
NP_103292
BAB49078
CAB94628
H83384
AAL46181
NP_396231
AAK90870
NP_109474
AAL19286
0.1

Figure 2

a)


Mus musculus
α1
α2
α3
α4
α5
Aspergillus niger
Agrobacterium radiobacter

b)

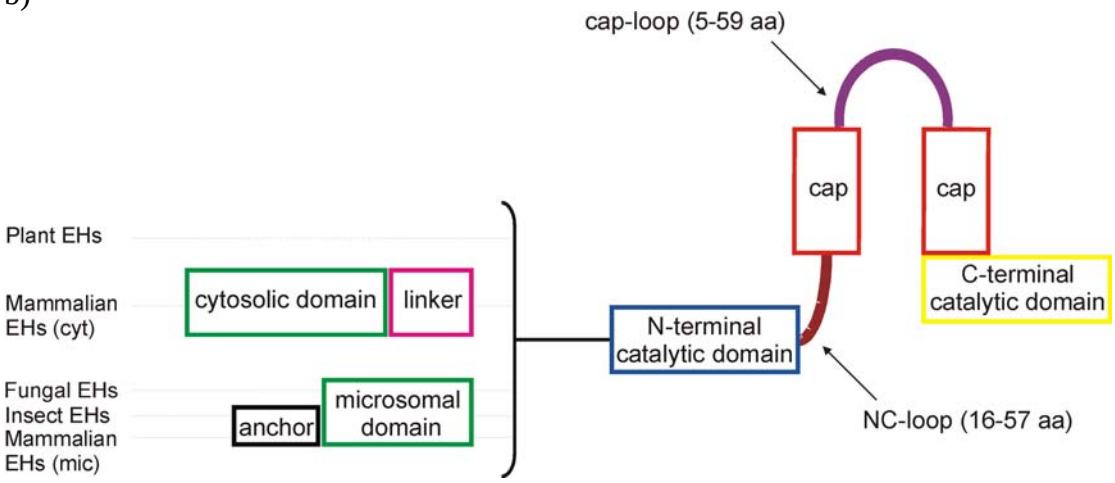

cap-loop (5-59 aa)
cap
cap
C-terminal catalytic domain
Plant EHs
Mammalian EHs (cyt)
cytosolic domain
linker
N-terminal catalytic domain
Fungal EHs
Insect EHs
Mammalian EHs (mic)
anchor
microsomal domain
NC-loop (16-57 aa)

Figure 3

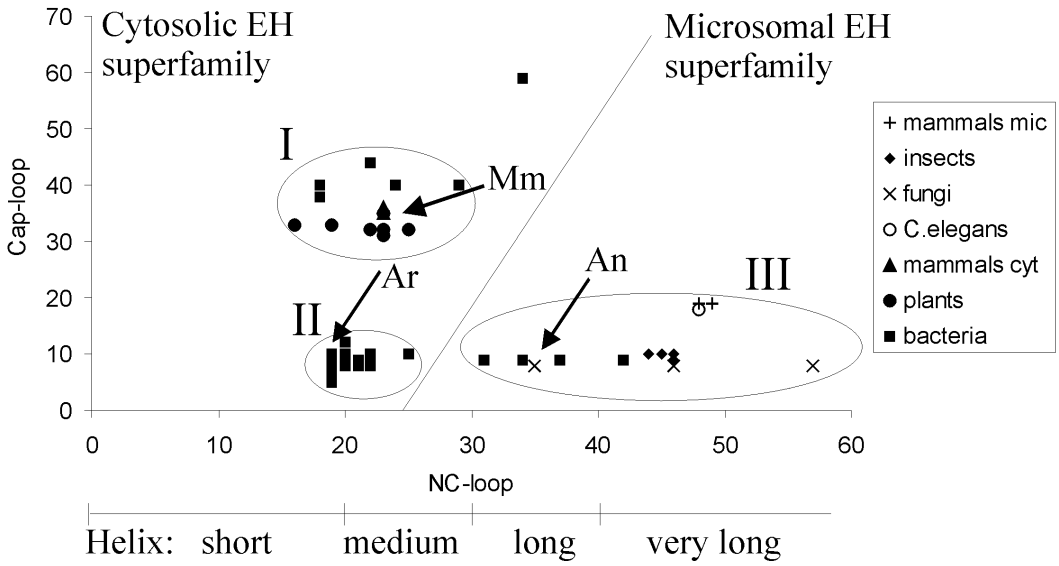

Cytosolic EH
superfamily
Microsomal EH
superfamily
I
II
III
Mm
Ar
An
Cap-loop
NC-loop
0
10
20
30
40
50
60
70
+ mammals mic
◆ insects
× fungi
○ C.elegans
▲ mammals cyt
● plants
■ bacteria
Helix: short medium long very long

Figure 4