Research

The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases

L Aravind and Eugene V Koonin

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

Correspondence: L Aravind. E-mail: aravind@ncbi.nlm.nih.gov

Published: 19 February 2001

Genome ***Biology*** 2001, **2(3)**:research0007.1–0007.8

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/3/research/0007

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 17 November 2000
Revised: 14 December 2000
Accepted: 12 January 2001

Abstract

Background: Protein fold recognition using sequence profile searches frequently allows prediction of the structure and biochemical mechanisms of proteins with an important biological function but unknown biochemical activity. Here we describe such predictions resulting from an analysis of the 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenases, a class of enzymes that are widespread in eukaryotes and bacteria and catalyze a variety of reactions typically involving the oxidation of an organic substrate using a dioxygen molecule.

Results: We employ sequence profile analysis to show that the DNA repair protein AlkB, the extracellular matrix protein leprecan, the disease-resistance-related protein EGL-9 and several uncharacterized proteins define novel families of enzymes of the 2OG-Fe(II) oxygenase superfamily. The identification of AlkB as a member of the 2OG-Fe(II) oxygenase superfamily suggests that this protein catalyzes oxidative detoxification of alkylated bases. More distant homologs of AlkB were detected in eukaryotes and in plant RNA viruses, leading to the hypothesis that these proteins might be involved in RNA demethylation. The EGL-9 protein from *Caenorhabditis elegans* is necessary for normal muscle function and its inactivation results in resistance against paralysis induced by the *Pseudomonas aeruginosa* toxin. EGL-9 and leprecan are predicted to be novel protein hydroxylases that might be involved in the generation of substrates for protein glycosylation.

Conclusions: Here, using sequence profile searches, we show that several previously undetected protein families contain 2OG-Fe(II) oxygenase fold. This allows us to predict the catalytic activity for a wide range of biologically important, but biochemically uncharacterized proteins from eukaryotes and bacteria.

Background

2-Oxoglutarate (2OG)- and Fe(II)-dependent dioxygenases are widespread in eukaryotes and bacteria and catalyze a variety of reactions typically involving the oxidation of an organic substrate using a dioxygen molecule [1,2]. One well-studied reaction catalyzed by such enzymes is the hydroxylation of proline and lysine sidechains in collagen and other animal glycoproteins [3-5]. In plants, enzymes of this family catalyze the formation of the plant hormone ethylene by oxidative desaturation of 1-aminocyclopropane-1-carboxylate, and catalyze the hydroxylation and desaturation steps in the synthesis of other plant hormones, pigments and metabolites such as gibberellins, anthocyanidins and flavones [1,6,7]. In bacteria and fungi, several members of this family participate in the desaturative cyclization and oxidative ring expansion reactions in the biosynthesis of antibiotics such as

penicillin and cephalosporin [8-10]. The details of the catalytic mechanism of these enzymes have been revealed by determination of the crystal structures of isopenicillin N synthase (IPNS), deacetoxycephalosporin C synthase (DAOCS) and clavaminic acid synthase (CAS) [8-11]. These structures showed that the catalytic core of the proteins consists of a double-stranded β-helix (DSBH) fold containing a HX[DE] dyad (where X is any amino acid) and a conserved carboxy-terminal histidine which together chelate a single iron atom. The substrates are bound within a spacious cavity formed by the interior of the DSBH (see the Structural Classification of Proteins (SCOP) [12]).

We use sequence profile analysis [13,14] to show that the DNA-repair protein AlkB, the extracellular matrix protein leprecan and the disease-resistance-related protein EGL-9 define new families of the 2OG-Fe(II) dioxygenase superfamily. AlkB is widely represented in bacteria and eukaryotes and has an important role in countering the toxic DNA modifications caused by alkylating agents in both *Escherichia coli* and *Homo sapiens* [15-17]. Despite considerable effort, the precise biochemical mechanisms of its action in DNA repair remain unknown. Recent studies have shown that AlkB is required for specifically processing lesions resulting from the alkylation of single-stranded (ss) DNA [18]. Our findings predict an unusual role for this enzyme in oxidative detoxification of DNA damage. The EGL-9 protein from *Caenorhabditis elegans* is necessary for normal muscle function, and its inactivation results in strong resistance to paralysis induced by the *Pseudomonas aeruginosa* toxin [19]. We predict that EGL-9 is a novel hydroxylase that could elicit its action through the modification of sidechains of intracellular proteins. Similarly, we show that the animal extracellular matrix protein leprecan [20] defines a hitherto unknown family of protein hydroxylases that might be involved in the generation of substrates for protein glycosylation.

Results and discussion

The 2OG-Fe(II) dioxygenase protein superfamily: classification and functional prediction

The Non-redundant Protein Sequence Database (NCBI) [21] was searched using the PSI-BLAST program [22] run to convergence, with a profile-inclusion threshold of 0.01 and AlkB protein sequences from various organisms as queries. In addition to the AlkB orthologs, these searches retrieved from the database, with statistically significant expectation (e) values, several other more distant homologs of AlkB, including uncharacterized eukaryotic proteins and fragments of the polyproteins of plant RNA viruses from the carla-, tricho- and potexvirus families. Examples of homologs found include: *Leishmania* L3377.4, iteration 5, e-value = 8×10^{-7}; *Drosophila* CG17807, iteration 3, e-value = 4×10^{-6}; papaya mosaic virus, iteration 3, e-value = 2×10^{-4}. Further iterations of the search using each of the detected proteins as a new query resulted in the detection of several more eukaryotic proteins, including EGL-9 and leprecan, several uncharacterized bacterial proteins and prolyl and lysyl hydroxylases. Finally, another iteration of database searches initiated with the sequences of bacterial proteins, typified by *E. coli* YbiX, resulted in the unification of these proteins with plant dioxygenases such as leucoanthocyanidin oxidase and gibberellin-20 oxidase. In this context, it should be noted that the DNA-repair proteins typified by *E. coli* AlkB are unrelated to the alkane omega-hydroxylase typified by the

Figure 1 *(see pages 3 and 4)*
Multiple sequence alignment of the 2OG-Fe(II) dioxygenase superfamily. Individual protein families are separated by blank lines and a brief description of each family is given to the right of the alignment. The numbers at the ends of the alignment indicate the position of the first and last of the aligned residues in the respective protein sequences. The consensus secondary structure is shown above the alignment in uppercase letters. It was derived by taking those elements that are shared by the predicted structures of individual families and the experimentally determined structures; H indicates α helix and E indicates extended conformation (β strand). The lowercase letters represent extensions of the secondary structure elements that are seen in some, but not all, members of the superfamily. The conserved amino-terminal extensions that are specific only to a given family are separated from the rest of the alignment by vertical lines. The coloring of the alignment columns is according to the 85% consensus that is shown underneath the alignment and includes the following categories of amino acid residues: h, hydrophobic; l, aliphatic; a, aromatic (Y, F, W, H, L, I, V, M, A, all shaded yellow); s, small (S, A, G, T, V, P, N, H, D, shaded blue); b, big (K, R, E, Q, W, F, Y, L, M, I, shaded gray); +, positively charged (K, R, H; colored magenta). The (predicted) catalytic residues are indicated by asterisks and with reverse red shading. The proteins are designated by the protein/gene name, the species abbreviation and the gene identification (GI) number. Protein abbreviations are: CAS, clavaminic acid synthase; DAOCS, deacetoxycephalosporin C synthetase; EFE, ethylene-forming enzyme; FLAS, flavonol synthase; Ga20Ox, giberellin 20-oxidase; IPNS, isopenicillin N synthase; LDOX, leucoanthocyanidin hydroxylase; Lep, leprecan; P4HA, prolyl-4-hydroxylase; PLO, lysyl hydroxylase; SanF and SanC, enzymes involved in nikkomycin biosynthesis. The remaining names are the standard names of the genes that encode the respective proteins. Species abbreviations: At, *Arabidopsis thaliana*; Bb, *Borrelia burgdorferi*; Cc, *Caulobacter crescentus*; Ce, *Caenorhabditis elegans*; Ci, *Ciona intestinalis*; Dm, *Drosophila melanogaster*; Ec, *Escherichia coli*; Em, *Emericella nidulans*; Hs, *Homo sapiens*; Lc, *Lysobacter lactamgenus*; Le, *Lycopersicon esculentum*; Mtu, *Mycobacterium tuberculosis*; Nc, *Neurospora crassa*; Pa, *Pseudomonas aeruginosa*; Pet, *Petunia hybrida*; Rr, *Rattus rattus*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*; Sot, *Solanum tuberosum*; Scoe, *Streptomyces coelicolor*; Scan, *Streptomyces ansochromogenes*; Scla, *Streptomyces clavuligerus*; Ssp, *Synechocystis*; Vc, *Vibrio cholerae*; ASPV, apple stem pitting virus; ACLSV, apple chlorotic leaf spot virus; BSV, blueberry scorch virus; GLV, garlic latent virus; GVA, grapevine virus A; PBCV, *Paramecium bursaria* chlorella virus; PMV, papaya mosaic virus; SHVX, shallot virus X.

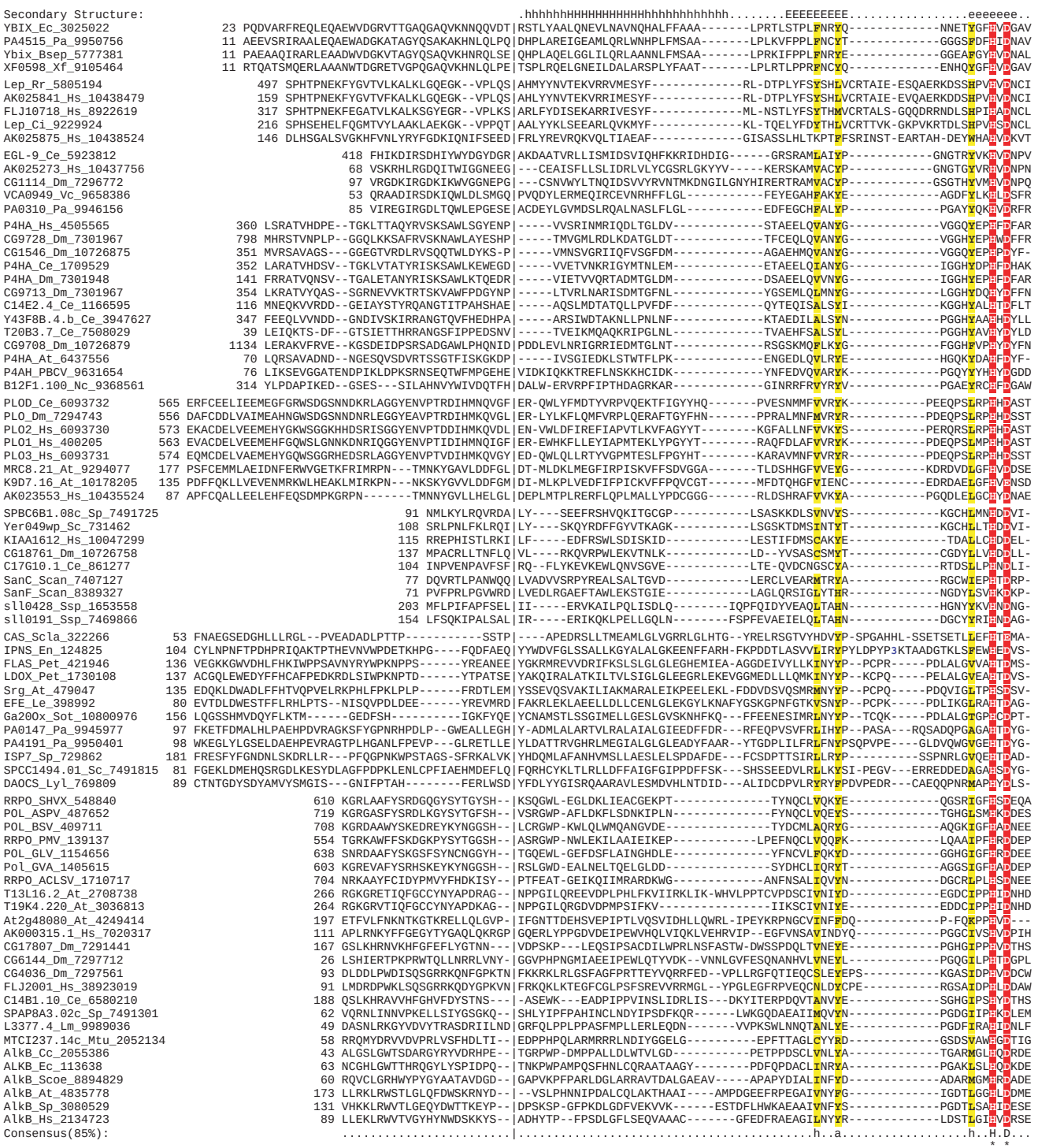

Figure 1 *(continued on the next page)*

Ps. oleovorans protein also named AlkB. Fortuitously, these latter alkane hydroxylases are also oxygenases; however, they are not 2OG-Fe(II) dioxygenases but a distinct class of di-iron enzymes [23]. On the basis of the results of database searches with representative sequences, we delineated several distinct families within the 2OG-Fe(II) dioxygenase fold and constructed individual alignments for each using the ClustalW program [24]. Secondary structure was predicted for each family using the PHD [25] and PSI-PRED [26] programs. Using the secondary structure elements

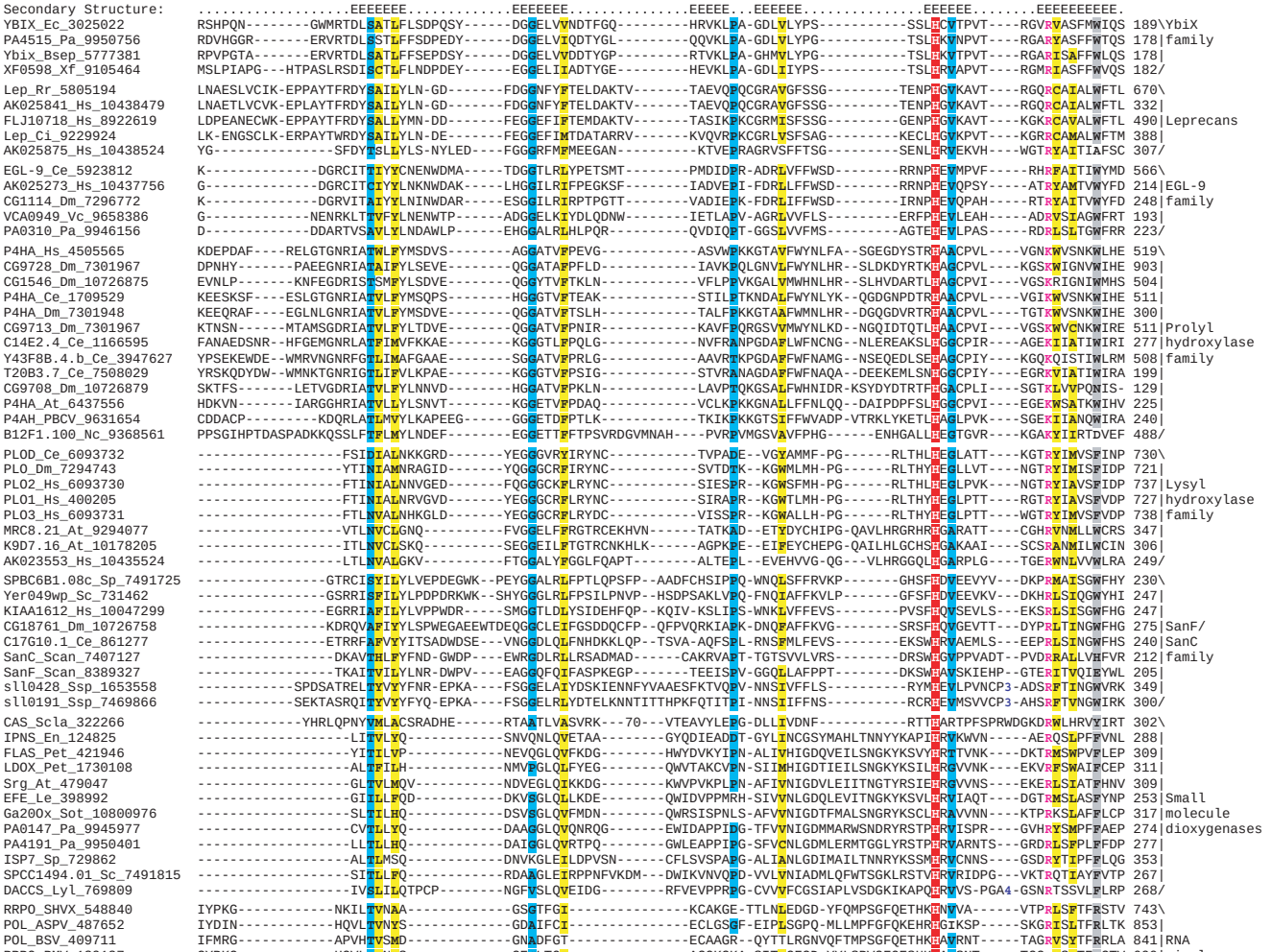

Figure 1 *(continued from the previous page)*

from the experimentally determined structures of IPNS (PDB:1ips), DAOCS (PDB:1dcs), CAS (PDB:1drt) and the predicted secondary structures for individual families, the conserved core of these elements was delineated (Figure 1). A multiple alignment for the entire superfamily was constructed by aligning the conserved sequence features shared by all the individual families. The boundaries of the secondary structure elements in the alignment were adjusted using the secondary structure conservation as a guide.

The conserved portion of the 2OG-Fe(II) dioxygenase superfamily proteins comprises the core DSBH domain seen in the IPNS, DAOCS and CAS structures [8-10] and part of the amino-terminal α helix that showed considerable variability

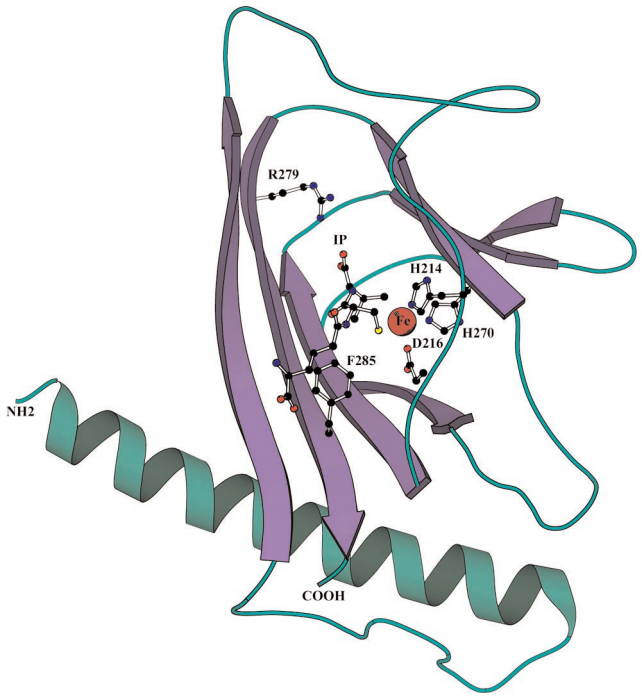

Figure 2
A structural model of the DSBH core of the 2OG-Fe(II) dioxygenase superfamily. This is based on the *Emericella nidulans* isopenicillin N synthase structure (PDB:1ips). The side chains of the amino acid residues implicated in catalysis and in substrate binding are shown (see text) and the Fe(II) ion is indicated by a red circle.

in both length and sequence between individual families (Figures 1,2). The DSBH region includes seven conserved strands that are common to all these proteins and are arranged in two sheets in a jelly-roll topology (Figures 2,3). However, different families have specific inserts in various positions between the conserved strands; some of these inserts contain additional secondary structures and show significant sequence conservation (Figures 1,3). For example, the insert between the fifth and sixth strand in AlkB is predicted to contain an extra strand, whereas in the small-molecule dioxygenase (IPNS/ethylene-forming enzyme (EFE)) family, the same region forms (or is predicted to form) a short helix (Figures 1,3). The clavaminic acid synthase, an outlier of this latter family, has its own characteristic inserts, including a giant (approximately 70 amino acids) insert between strands 4 and 5 [8], and some members of the AlkB family have smaller inserts in the same position (Figure 1). This reflects the relative resilience of the core DSBH to insertions, and accounts for difficulties in unification of this superfamily by sequence-based methods. The multiple alignment contains at least three characteristic conserved motifs that center, respectively, at a HXD dyad near the amino terminus, a histidine towards the carboxyl terminus, and an arginine or lysine further downstream (Figure 1). The HXD dyad is located in a flexible loop that follows the first conserved strand and stacks with the sheet containing three of the core strands (Figures 2,3). The second conserved histidine is associated with the beginning of the sixth strand, whereas the conserved basic residue (R or K) is in the beginning of the seventh strand of the DSBH core (Figures 2,3). The sixth position after the conserved basic residue is invariably occupied by a bulky residue that is either arginine (in AlkB) or phenylalanine or tryptophan in all other members of this fold [8-10] (Figure 1).

The roles of all these conserved residues in catalysis are apparent from the crystal structures of IPNS, clavaminic acid synthase (CAS) and deacetoxycephalosporin C synthase (DAOCS) [8-10]. The conserved HXD and the carboxy-terminal histidine coordinate Fe(II) that is directly involved in catalysis by these enzymes (Figure 2). The conserved basic residue interacts with the carboxylate group of the acidic substrate, while the bulky or aromatic residue located carboxy-terminal to the basic one forms the base of the cleft that holds this substrate molecule (Figure 2). Whereas most of these enzymes necessarily use 2-OG as the acidic substrate, IPNS does not bind 2-OG. Its place is occupied by its single substrate, L-δ-(α-aminoadipoyl)-L-cysteinyl-D-valine, whose carboxyl sidechain interacts with the conserved basic residue similarly to the OG-carboxylate in the other enzymes of this superfamily [9,10].

The conservation of all the residues implicated in catalysis in the biochemically uncharacterized proteins such as AlkB, EGL-9, leprecan and YbiX suggests they all catalyze oxidative reactions similar to those catalyzed by IPNS, DAOCS, CAS and related enzymes such as protein lysyl/prolyl hydroxylases, EFE and leucoanthocyanidin oxidases [1,2]. Using the available contextual information, we attempted to predict possible substrates of these proteins. AlkB binds ssDNA and is required for the processing of toxic DNA modifications caused by SN2 alkylating agents such as methylmethanesulfonate (MMS) specifically on ssDNA. On the basis of the preferential specificity of AlkB-dependent repair for ssDNA and for SN2 as opposed to SN1 alkylating agents, it has been proposed that its targets include modified bases such as N^1-methyladenine and N^3-methylcytosine [18]. The predicted 2OG-Fe(II) dioxygenase activity of AlkB is probably involved in the detoxification of methylated bases in ssDNA, possibly through hydroxylation of the methyl groups, resulting in less toxic base derivatives. The hydroxylation might be followed by a second oxidative step that could remove the hydroxymethyl group, restoring the normal base in DNA. These reactions are consistent with the observation that, unlike AlkA, AlkB has no DNA glycosylase activity [18]. The most intriguing finding made during our analysis of the AlkB family is the existence of multiple eukaryotic AlkB homologs, other than the actual orthologs, especially in plants and their viruses. The specific action of AlkB on ssDNA suggests that ssRNA could be the substrate

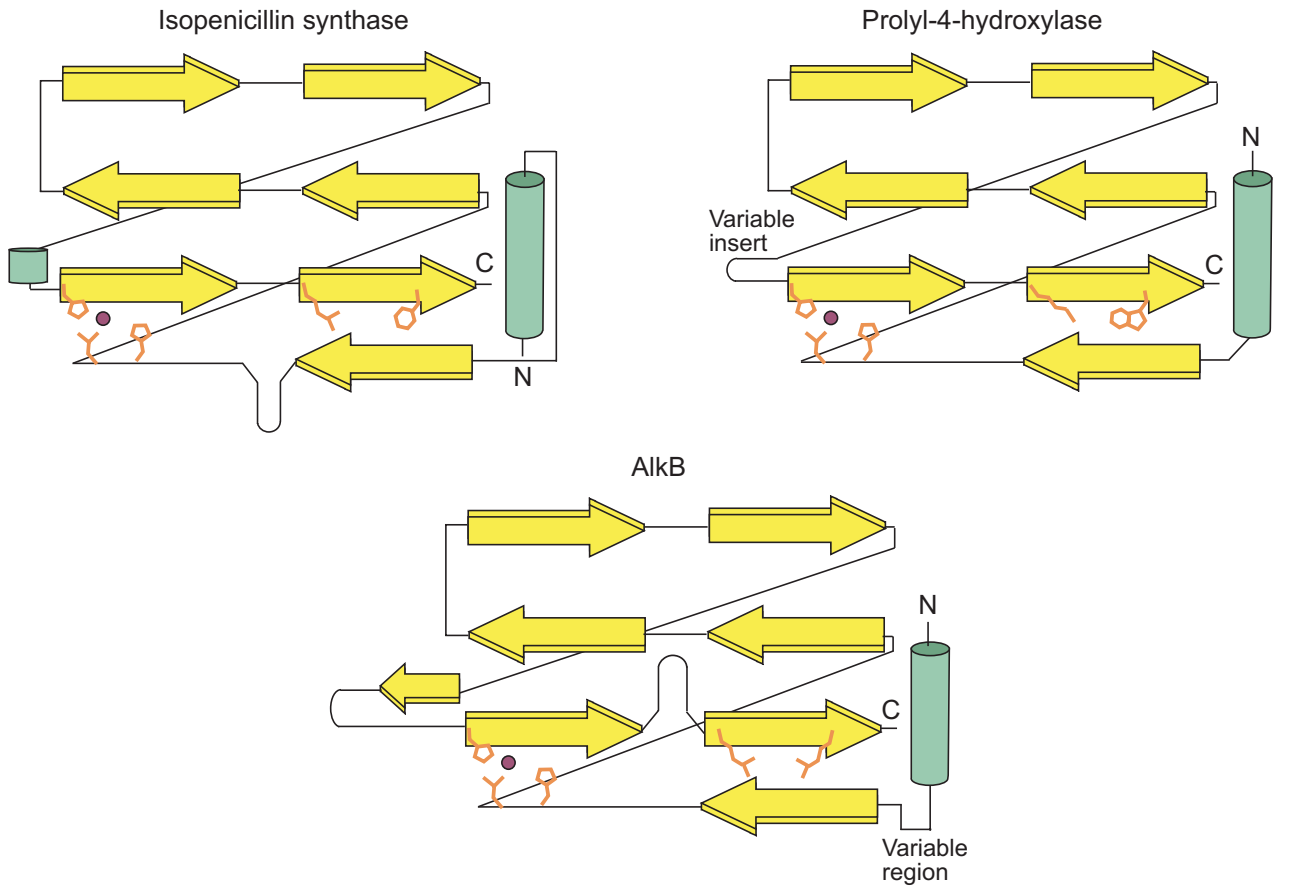

Figure 3
Topological diagrams for three members of the 2OG-Fe(II) dioxygenase superfamily. The diagrams are based on the experimentally determined structures for *E. nidulans* isopenicillin N synthase (PDB: 1ips) and structural models of prolyl-4-hydroxylase and AlkB. The amino acid residues of the active site and the Fe(II) ion are shown as in Figure 2.

for other members of this family. Given the presence of RNA methylases that could modify RNA in a similar way as the alkylating agents the function of the uncharacterized AlkB-like proteins could be to reverse such modifications. Such modifications might also be used in the host-mediated inactivation of viral RNAs, and the AlkB homologs acquired by some plant viruses could counter this host-defense mechanism. The connection between nucleic acid methylation and the AlkB homologs described here is supported by the domain architecture of a conserved pair of animal proteins (C14B1.10 from *C. elegans* and CG17807 from *Drosophila*) in which an amino-terminal AlkB-like domain is fused to a carboxy-terminal predicted methyltransferase domain. These proteins could potentially be involved in regulation of RNA stability or DNA repair via controlled methylation/demethylation.

EGL-9 defines a new family within the 2OG-Fe(II) dioxygenase superfamily that is highly conserved in animals and is also represented in the pathogenic proteobacteria such as *Ps. aeruginosa* and *Vibrio cholerae*. In *C. elegans*, EGL-9 is required for normal egg laying, whereas loss of its function provides resistance to hypercontractile muscular paralysis caused by *Ps. aeruginosa* [19]. The vertebrate homolog of EGL-9 is specifically expressed in smooth muscles and is likely to have a role in their function [27]. At its carboxyl terminus, EGL-9 contains a MYND finger domain that is probably involved in specific protein-protein interactions [28]. The closest relatives of the EGL-9 family are the proline hydroxylases with which they share a region of specific extended conservation amino terminal to the core DSBH domain (Figure 1). This relationship, along with the combination to the intracellular MYND domain and the lack of signal peptides, suggests that the Egl-9 family proteins are prolyl hydroxylases that modify intracellular proteins, unlike the classic prolyl hydroxylases that have been implicated primarily in the modification of collagens in the endoplasmic lumen. The interesting aspect of the EGL-9 family is its presence in *V. cholerae* and *Ps. aeruginosa* (Figure 1), which have apparently acquired these genes by horizontal transfer

from eukaryotes. The direct connection between this gene acquisition and the action of the *Ps. aeruginosa* toxins on animal muscles is unclear, but it seems possible that the bacterial EGL-9-like proteins modify host proteins in a manner that favors the survival and spread of the pathogen. This might be especially pertinent if the host downregulates the endogenous ortholog in response to the infection.

Leprecan is a proteoglycan that is associated with the basement membrane in chordates [20]. It and related proteins contain an amino-terminal segment rich in leucine and proline [20] and a carboxy-terminal globular part that includes the 2OG-Fe(II) oxygenase domain. More distant relatives of the leprecan-like proteins include the T23K23.7 protein from *Arabidopsis* and a family of uncharacterized proteobacterial proteins typified by YbiX (Figure 1). These proteins are predicted to be previously unnoticed amino-acid hydroxylases that catalyze modifications of intracellular and extracellular proteins.

Evolutionary implications

Sequence conservation is high within individual families of the 2OG-Fe(II) dioxygenase superfamily, with specific extensions typical of each family, but low between different families (Figure 1). This observation, together with the phyletic distribution of these proteins, provides some clues to their evolutionary history. Members of this superfamily could not be detected in the Archaea despite extensive profile searches of the archaeal proteomes as well as transitive searches seeded with many divergent sequences as starting points. This suggests that, unlike bacteria and eukaryotes, in which the superfamily is widely represented, archaea do not encode *bona fide* members. A corollary of this is that horizontal gene transfer between bacteria and eukaryotes might have had a significant role in the evolution of this superfamily. The DSBH fold that comprises the core domain of the 2OG-Fe(II) dioxygenases is also present in a large number of proteins typified by the arabinose-binding domain of the transcription regulator AraC, plant seed proteins such as vicilin, and oxalate oxidase ([29]; see SCOP [12]). Despite the lack of detectable sequence similarity to 2OG-Fe(II) dioxygenases, these proteins contain a conserved HXH motif and a carboxy-terminal histidine that appear to be equivalent to the metal-chelating HXD and carboxy-terminal histidine of the latter (see above). Thus, these two protein superfamilies could have evolved from a common ancestor by acquiring distinct catalytic and ligand-binding properties. The classic AraC-like DSBH proteins appear to have a more universal distribution than the 2OG-Fe(II) dioxygenases, with diverse forms represented in archaea, bacteria and eukaryotes. The 2OG-Fe(II) dioxygenase superfamily is present in diverse bacteria that had probably diverged before the diversification of the eukaryotes that contain this superfamily. This, together with the phyletic distributution of the AraC and 2OG-Fe(II) dioxygenase superfamilies, leads us to speculate that the 2OG-Fe(II) dioxygenase superfamily evolved from the AraC-like superfamily in bacteria through drastic sequence divergence that eliminated significant sequence similarity, followed by fixation of the modified active-site configuration.

In the case of the AlkB family, a single horizontal transfer event probably resulted in their entry into the eukaryotic lineage, which was followed by adaptation to new, RNA-related roles by some paralogs. In the case of the small-molecule dioxygenase (IPNS/EFE) family, a complex web of relationships can be discerned. The EFE and the plant secondary metabolite biosynthesis enzymes flavonol synthase, leucoanthocyanidin hydroxylase and giberellin-20 oxidase show maximum diversity in the plant lineage [1,6,7]. Their close homologs are, however, also found in *Pseudomonas*, but not in other well-studied proteobacterial lineages. Similarly, fungi contain several enzymes involved in secondary metabolite biosynthesis, such as IPNS and DAOCS, that are distinctly related to their counterparts in actinomycetes. This patchy phyletic distribution across kingdoms is suggestive of multiple gene transfer events that have apparently led to the wide dissemination of these proteins in bacteria and eukaryotes. The close grouping of the ethylene-forming enzyme and its *Pseudomonas* homologs [30] to the exclusion of other members of this family in plants indicates a possible recent acquisition of the gene in *Pseudomonas* from its plant hosts. In *Arabidopsis* there is an expansion of small-molecule dioxygenases (at least 75 members), whereas *Ps. aeruginosa* has at least three recently duplicated members. This proliferation of the small-molecule dioxygenases is consistent with their possible role in the synthesis of secondary metabolites in these organisms [1,6,7,30]. The amino-acid hydroxylases show the greatest diversity in eukaryotes, and are represented in a number of different forms in both animals and plants. They were probably derived from small-molecule hydroxylases early in eukaryotic evolution. In contrast, the predicted amino-acid hydroxylases of the EGL-9 family are seen sporadically, in single copies, in certain bacterial lineages such as *V. cholerae* and *Ps. aeruginosa,* suggesting a secondary horizontal transfer from the eukaryotes to bacteria.

Conclusions

Before this study, structure determination, biochemical studies and sequence comparisons of 2OG-Fe(II) dioxygenases [1,9] had elucidated their structural fold, active-site residues and reaction mechanism. Here, using sequence profile searches, we show that many other protein families contain the same constellation of active-site residues and are predicted to adopt the same fold. This allows us to predict the catalytic activity of a wide range of functionally important, but biochemically uncharacterized, proteins from eukaryotes and bacteria. In particular, we propose a specific mechanism of action in DNA repair, and possibly in RNA modification, for the AlkB protein and its homologs.

Materials and methods

The Non-redundant Protein Sequence Database [21], the Expressed Sequence Tags Database (NCBI) [31] and the individual protein sequence databases of completely and partially sequenced genomes [32] were searched using the gapped version of the BLAST programs (BLASTPGP for proteins and TBLASTNGP for translating searches of nucleotide databases) [22]. Sequence profile searches were performed using the PSI-BLAST program [22]; profiles were saved using the –C option and retrieved using the –R option. Multiple alignments of amino acid sequences were generated using a combination of PSI-BLAST, CLUSTALW [24] and secondary structure predictions that were produced using the PHD program [25] and the PSI-PRED program [26], with multiple alignments of individual protein families used as queries. The three-dimensional structure visualization, alignment and modeling were carried out using the SWISS-PDB-Viewer program [33].

References

1. Prescott AG: **A dilemma of dioxygenases: or where molecular biology and biochemistry fail to meet.** *J Exp Bot* 1993, **44:**849-861.
2. Hegg EL, Que L Jr: **The 2-His-1-carboxylate facial triad - an emerging structural motif in mononuclear non-heme iron(II) enzymes.** *Eur J Biochem* 1997, **250:**625-629.
3. Myllyharju J, Kivirikko KI: **Characterization of the iron- and 2-oxoglutarate-binding sites of human prolyl 4-hydroxylase.** *EMBO J* 1997, **16:**1173-1180.
4. Pirskanen A, Kaimio AM, Myllyla R, Kivirikko KI: **Site-directed mutagenesis of human lysyl hydroxylase expressed in insect cells. Identification of histidine residues and an aspartic acid residue critical for catalytic activity.** *J Biol Chem* 1996, **271:**9398-9402.
5. Passoja K, Myllyharju J, Pirskanen A, Kivirikko KI: **Identification of arginine-700 as the residue that binds the C-5 carboxyl group of 2-oxoglutarate in human lysyl hydroxylase 1.** *FEBS Lett* 1998, **434:**145-148.
6. Zhang Z, Barlow JN, Baldwin JE, Schofield CJ: **Metal-catalyzed oxidation and mutagenesis studies on the iron(II) binding site of 1-aminocyclopropane-1-carboxylate oxidase.** *Biochemistry* 1997, **36:**15999-16007.
7. Lukacin R, Britsch L: **Identification of strictly conserved histidine and arginine residues as part of the active site in *Petunia hybrida* flavanone 3beta-hydroxylase.** *Eur J Biochem* 1997, **249:**748-757.
8. Zhang Z, Ren J, Stammers DK, Baldwin JE, Harlos K, Schofield CJ: **Structural origins of the selectivity of the trifunctional oxygenase clavaminic acid synthase.** *Nat Struct Biol* 2000, **7:**127-133.
9. Valegard K, van Scheltinga AC, Lloyd MD, Hara T, Ramaswamy S, Perrakis A, Thompson A, Lee HJ, Baldwin JE, Schofield CJ, *et al.*: **Structure of a cephalosporin synthase.** *Nature* 1998, **394:**805-809.
10. Roach PL, Clifton IJ, Fulop V, Harlos K, Barton GJ, Hajdu J, Andersson I, Schofield CJ, Baldwin JE: **Crystal structure of isopenicillin N synthase is the first from a new structural family of enzymes.** *Nature* 1995, **375:**700-704.
11. Lange SJ, Que L Jr: **Oxygen activating nonheme iron enzymes.** *Curr Opin Chem Biol* 1998, **2:**159-172.
12. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C: **SCOP: a structural classification of proteins database.** *Nucleic Acids Res* 2000, **28:**257-259.
13. Altschul SF, Koonin EV: **Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases.** *Trends Biochem Sci* 1998, **23:**444-447.
14. Aravind L, Koonin EV: **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.** *J Mol Biol* 1999, **287:**1023-1040.
15. Wei YF, Carter KC, Wang RP, Shell BK: **Molecular cloning and functional analysis of a human cDNA encoding an *Escherichia coli* AlkB homolog, a protein involved in DNA alkylation damage repair.** *Nucleic Acids Res* 1996, **24:**931-937.
16. Chen BJ, Carroll P, Samson L: **The *Escherichia coli* AlkB protein protects human cells against alkylation-induced toxicity.** *J Bacteriol* 1994, **176:**6255-6261.
17. Kondo H, Nakabeppu Y, Kataoka H, Kuhara S, Kawabata S, Sekiguchi M: **Structure and expression of the *alkB* gene of *Escherichia coli* related to the repair of alkylated DNA.** *J Biol Chem* 1986, **261:**15772-15777.
18. Dinglay S, Trewick SC, Lindahl T, Sedgwick B: **Defective processing of methylated single-stranded DNA by *E. coli* AlkB mutants.** *Genes Dev* 2000, **14:**2097-2105.
19. Darby C, Cosma CL, Thomas JH, Manoil C: **Lethal paralysis of *Caenorhabditis elegans* by *Pseudomonas aeruginosa*.** Proc Natl Acad Sci USA 1999, **96:**15202-15207.
20. Wassenhove-McCarthy DJ, McCarthy KJ: **Molecular characterization of a novel basement membrane-associated proteoglycan, leprecan.** J Biol Chem 1999, **274:**25004-25017.
21. **Non-redundant Protein Sequence Database** [http://www.ncbi.nlm.nih.gov/BLAST/]
22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** Nucleic Acids Res 1997, **25:**3389-3402.
23. Shanklin J, Achim C, Schmidt H, Fox BG, Munck E: **Mossbauer studies of alkane omega-hydroxylase: evidence for a diiron cluster in an integral-membrane enzyme.** *Proc Natl Acad Sci USA* 1997, **94:**2981-2986.
24. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
25. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232:**584-599.
26. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol* 1999, **292:**195-202.
27. Wax SD, Rosenfield CL, Taubman MB: **Identification of a novel growth factor-responsive gene in vascular smooth muscle cells.** *J Biol Chem* 1994, **269:**13041-13047.
28. Masselink H, Bernards R: **The adenovirus E1A binding protein BS69 is a corepressor of transcription through recruitment of N-CoR.** *Oncogene* 2000, **19:**1538-1546.
29. Gane PJ, Dunwell JM, Warwicker J: **Modeling based on the structure of vicilins predicts a histidine cluster in the active site of oxalate oxidase.** *J Mol Evol* 1998, **46:**488-493.
30. Nagahama K, Yoshino K, Matsuoka M, Sato M, Tanase S, Ogawa T, Fukuda H: **Ethylene production by strains of the plant-pathogenic bacterium *Pseudomonas syringae* depends upon the presence of indigenous plasmids carrying homologous genes for the ethylene-forming enzyme.** *Microbiology* 1994, **140:**2309-2313.
31. **Expressed Sequence Tags Database** [http://www.ncbi.nlm.nih.gov/blast/blast.cgi?Jform=0]
32. **Unfinished Genomes Database** [http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html]
33. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723.