Protein Science (1999), *8*:137–146. Cambridge University Press. Printed in the USA.

Relationships within the aldehyde dehydrogenase extended family

JOHN PEROZICH,[1] HUGH NICHOLAS,[2] BI-CHENG WANG,[3] RONALD LINDAHL,[4]
AND JOHN HEMPEL[5]

[1]Department of Molecular Genetics and Biochemistry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15261
[2]Pittsburgh Supercomputing Center, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213
[3]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia 30602
[4]Department of Biochemistry and Molecular Biology, University of South Dakota, Vermillion, South Dakota 57069
[5]Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received April 7, 1998; Accepted September 13, 1998)

Abstract

One hundred-forty-five full-length aldehyde dehydrogenase-related sequences were aligned to determine relationships within the aldehyde dehydrogenase (ALDH) extended family. The alignment reveals only four invariant residues: two glycines, a phenylalanine involved in NAD binding, and a glutamic acid that coordinates the nicotinamide ribose in certain E-NAD binary complex crystal structures, but which may also serve as a general base for the catalytic reaction. The cysteine that provides the catalytic thiol and its closest neighbor in space, an asparagine residue, are conserved in all ALDHs with demonstrated dehydrogenase activity. Sixteen residues are conserved in at least 95% of the sequences; 12 of these cluster into seven sequence motifs conserved in almost all ALDHs. These motifs cluster around the active site of the enzyme. Phylogenetic analysis of these ALDHs indicates at least 13 ALDH families, most of which have previously been identified but not grouped separately by alignment. ALDHs cluster into two main trunks of the phylogenetic tree. The largest, the "Class 3" trunk, contains mostly substrate-specific ALDH families, as well as the class 3 ALDH family itself. The other trunk, the "Class 1/2" trunk, contains mostly variable substrate ALDH families, including the class 1 and 2 ALDH families. Divergence of the substrate-specific ALDHs occurred earlier than the division between ALDHs with broad substrate specificities. A site on the World Wide Web has also been devoted to this alignment project.

Keywords: aldehyde dehydrogenase (ec 1.2.1.3); multiple sequence alignment; protein family

Aldehyde dehydrogenases (ALDHs) catalyze the oxidation of aldehydes to their corresponding carboxylic acids and occur throughout all phyla. Many disparate aldehydes are ubiquitous in nature and most are toxic at low levels because of their chemical reactivity. Thus, levels of metabolic-intermediate aldehydes must be carefully regulated. For this, most well-studied organisms are known to have several distinct ALDHs, which take part in a variety of physiological roles. Some ALDHs are highly specific for a very limited range of substrates while others show a broad substrate specificity. All ALDHs require either NAD or NADP as a cofactor (reviewed, Lindahl, 1992; Yoshida et al., 1998).

Within a decade of the first ALDH sequence, alignment of 16 of the then most divergent ALDH sequences (Hempel et al., 1993) supported a common, conserved ALDH structure and suggested residues with important structural and functional roles, similar to findings in other enzyme families (Jörnvall, 1977; Brändén & Tooze, 1991; Creighton, 1993). Since then several novel examples of ALDHs, such as nonphosphorylating glyceraldehyde-3-phosphate dehydrogenases (Habenicht et al., 1994) and PutA fusion proteins (Ling et al., 1994), have been identified and numerous other ALDH sequences of various types reported.

Just recently the first two ALDH tertiary structures have been reported (Liu et al., 1997; Steinmetz et al., 1997). Since many forthcoming studies on ALDHs will depend on dissection of these molecular structures, it is useful to "take a step back" and examine the ALDH extended family as a whole, allowing information based on the known tertiary structures to be more readily be applied to other more diverse ALDHs. In addition new information on ALDH

Reprint requests to: John Hempel, 301 Clapp Hall, University of Pittsburgh, Pittsburgh, Pennsylvania 15260; e-mail: hempel@psc.edu.

Abbreviations: ALDH, aldehyde dehydrogenase; BALDH, betaine aldehyde dehydrogenase; DH, dehydrogenase; FTDH, 10-formyltetrahydrofolate dehydrogenase; SSALDH, succinic semialdehyde dehydrogenase; MMSALDH, methylmalonyl semialdehyde dehydrogenase; GGSALDH, γ-glutamyl semialdehyde dehydrogenase; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; HMSALDH, 2-hydroxymuconic semialdehyde dehydrogenase; GABA, γ-aminobutyric acid; PAM, accepted point mutations per 100 residues.

function and evolutionary origins may be discovered. In the current study, 145 available full-length ALDH sequences were aligned.

Results and discussion

All 145 full-length ALDH sequences available to us by September 1997 were compiled and aligned. Final adjustments were made manually using the Dayhoff PAM 250 scoring matrix. Several of the 145 sequences are from longer fusion proteins with domains of separate lineage joined to an ALDH domain (below). In these cases, we used only that sequence domain demonstrably homologous to other ALDHs. Most organisms have several distinct ALDH genes. The largest number of sequences in the current alignment from any one species is 13, from *Escherichia coli*. Eleven human sequences are present, excluding the ALDH8 protein (Hsu et al., 1997), which does not fit our criteria for a full-length ALDH (below). There are also 10 ALDHs from *Bacillus subtilis* and 14 from various species of *Pseudomonas*, most of which are involved with metabolizing aromatic aldehydes.

We have created a World Wide Web site devoted to ALDHs with the URL www.psc.edu/biomed/pages/research/Col_HBN_ALDH.html. This site will contain all figures, tables, and statistical data associated with this manuscript, along with others too large for publication. The identities and references for all sequences used in this study are available at this site. Names were assigned based on the source organism followed by the ALDH family to which the sequence may belong, based on the name given in the sequence submission. For presentation here, family consensus sequences were aligned (Fig. 1), as the actual alignment of 145 ALDH sequences (available at the ALDH web site) is too large for print format. Index position numbers in the consensus (Fig. 1) and complete alignments are identical.

Residue nomenclature

References to secondary structure use nomenclature from the rat cytosolic class 3 ALDH structure (Liu et al., 1997). Specific residues are identified by their position number in rat cytosolic class 3 ALDH (rat-3), followed by the index number (Fig. 1) in brackets.

Criteria for inclusion in the alignment

Typically, ALDH subunits range in size from about 450 to 520 amino acids. We have included as full-length sequences only those which include the region between Arg25 [index 166] and Gly414 [659]. This region includes all invariant residues and conserved segments identified previously in ALDHs (Hempel et al., 1993).

Problematic sequences

The alignment shown in Figure 1 contains gaps that are present in all of the consensus sequences, resulting from idiosyncratic insertions in certain individual sequences. Mtspn-beta has a short internal repeat (FEYFEY), which causes a three position gap in all other ALDHs from indices 216–218. No other known ALDH has a repeat or other insertion here. Celeg-2 has a large, unique insertion at indices 222–236. Celeg-FTDH contains a unique deletion among ALDHs from indices 300–335, which spans β-2 and α-B in the class 3 ALDH tertiary structure (Liu et al., 1997). The reported sequence of the PutA protein from *Bradyrhizobium* (Brady-PutA) has a region of divergent, low complexity sequence (indices 335–382) that spans the integral region involved in coenzyme binding (Liu et al., 1997). When we translated this *putA* gene in all three frames (unpubl. data), "frame a" yields the reported protein sequence. However, the "frame b" translation results in a protein sequence that very closely resembles other ALDHs between indices 335–382. It remains to be determined whether this represents an actual frameshift. We have used the published sequence for this alignment, although we are inclined to believe that there is a sequencing error as the large number of reported positively charged amino acids in this region would disrupt coenzyme binding.

Residue conservation

Only 4 of the 23 invariant residues identified in the previous alignment of 16 ALDHs (Hempel et al., 1993) remain so in the current alignment of 145 sequences: Gly187 [368], Gly240 [434], Glu333 [561], and Phe335 [563]. Gly187 [368] and Phe335 [563] are integral for binding the nicotinamide ring of NAD. By its extreme Ramachandran angles (Liu et al., 1997), Gly240 [434] appears necessary to position the catalytic nucleophile, Cys243 [437]. Glu333 [561] has been proposed to act in binding NAD (Ni et al., 1997; Steinmetz et al., 1997), although alternate evidence exists (Hempel et al., 1999).

Excluding the 4 invariant residues, 12 residues are found in more than 95% of the sequences: Arg25 [166], Gly105 [272], Asn114 [281], Pro116 [283], Gly131 [300], Lys137 [306], Gly211 [400], Cys243 [437], Pro337 [564], Gly383 [618], Asn388 [623], and Gly403 [644]. The glycines and prolines, which represent 9 of the 16 highly conserved and invariant residues, all lie at critical turns in the class 3 ALDH structure (Hempel et al., 1997). For example, Gly211 [400] is the first glycine of a Gly-Gly dipeptide that marks the boundary between the coenzyme-binding and catalytic domains. Also, Gly403 [644] is involved in the "U-turn" region (Liu et al., 1997). Glycines are overrepresented among the conserved residues in ALDHs. Similar observations have been made in other enzyme families, such as short- and long-chain alcohol dehydrogenases and "Rec-A like" proteins (Jörnvall, 1977; Persson et al., 1991; Brocchieri & Karlin, 1998).

Cys243 [437], the catalytic thiol, is present in all sequences with catalytic activity. Cephalopod Ω-crystallins have arginines present at this position, but have been reported to lack ALDH activity (Zinovieva et al., 1993). Arg25 [166] is conserved in all but two sequences, the Ω-crystallin from squid (Squid-oxt), which lacks activity, and one of two reported PutA protein sequences from *E. coli* (Ecoli-pro2). Only two ALDHs lack Lys137 [306], which has been proposed to hydrogen bond to the adenine ribose (Ni et al., 1997): Mejan-g3p and Ecoli-AldA. Asn388 [623] serves to terminate strand β-12.

Asn114 [281] has been proposed to stabilize the carbonyl oxygen of the substrate aldehyde during catalysis (Steinmetz et al., 1997; Hempel et al., 1999). Only two sequences from molds (Altal-aldh and Clahe-aldh) lack this residue, with Glu present instead. However, we note (1) the segments flanking this position in these sequences do not fit the motif that includes this residue (below; Table 1), and (2) these two ALDHs were identified specifically as allergens without any enzymatic ability having been reported (Achatz et al., 1995). We also note that the residue at this position in rat formyltetrahydrofolate dehydrogenase has been revised to Asn (Krupenko et al., 1997). As well, the originally-submitted Rat-ret

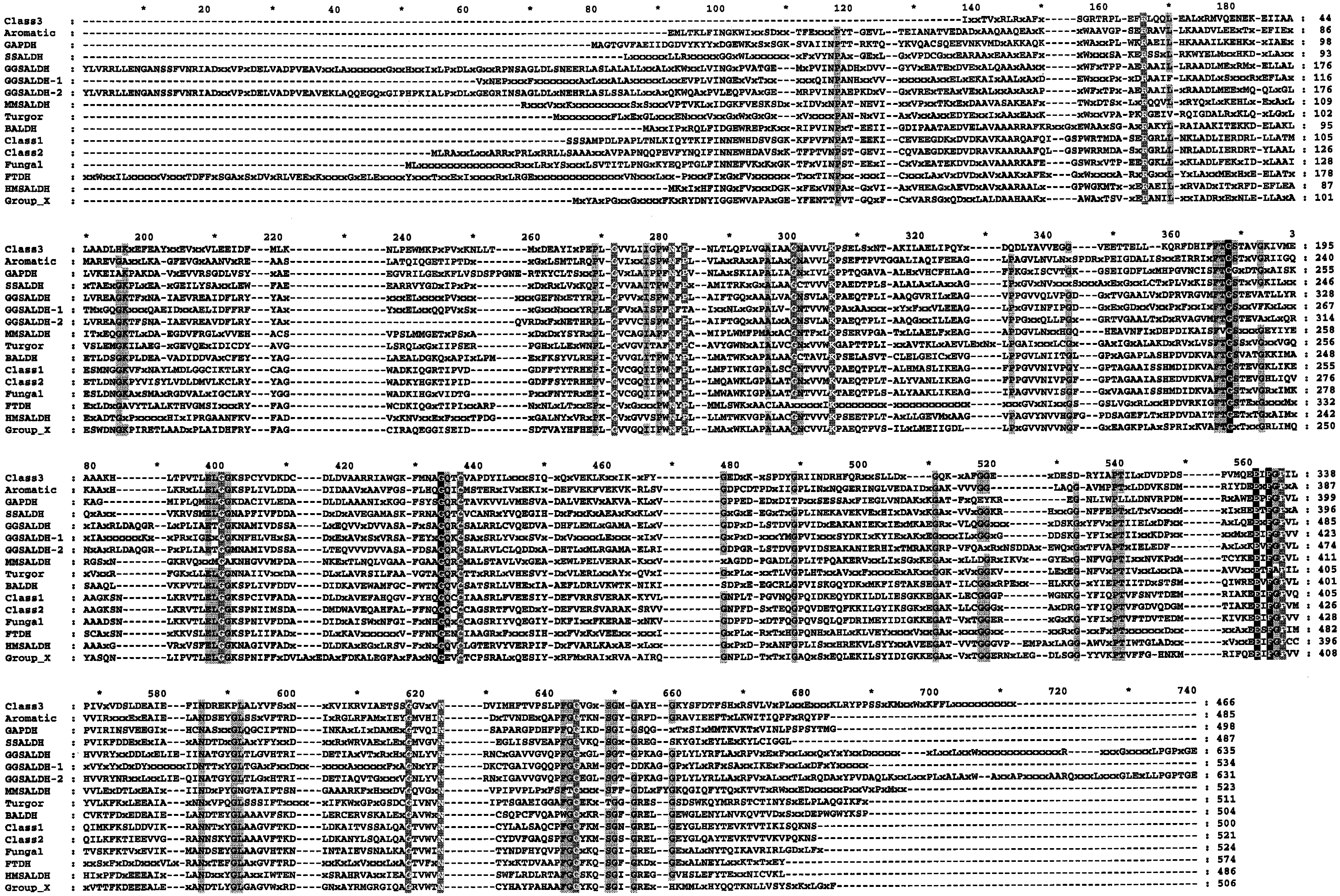

Fig. 1. Alignment of the consensus sequences of ALDH families. Invariant residues are highlighted in black. Residues that are at least 95% conserved are highlighted in gray with white letters. Residues conserved in 80% of the sequences are shaded gray with black letters. Residue positions are denoted by index numbers above the first sequence in the alignment. Index numbers are given at every 20th position, while every tenth position is denoted by an asterisk. These index numbers are identical to those in the complete alignment, available at the ALDH Web site. Sequence names are based on the assigned abbreviations. The GGSALDH sequence is the combined consensus for both type I (GGSALDH-1) and type II (GGSALDH-2) GGSALDHs. X's in sequences denote positions without clear residue consensus within that family.

Table 1. *Ten most conserved sequence motifs in ALDHs*

Motif number[a]	Length	Information content (bits)	Motif[b]	Indices
1	5	19.9	[Past]-[WFy]-[**N**e]-[FYgalv]-[**P**tl]	279–283
2	14	35.2	[Apnci]-[Liamv]-[Avslcimg]-[ACtlmvgf]-**G**-[Ncdi]-[Tavcspg]-[Vaimfcltgy]-[Vil]-[Lvmiwafhcy]-[**K**h]-[Ptvghms]-[ASdhp]-[Epsadqgilt]	296–309
3	10	22.4	[Grkpwhsay]-[FLeivqnarmhk]-[Pg]-[Plakdievsrf]-[Gnde]-[Vliat]-[VLifyac]-[Nglqshat]-[VIlyaqgfst]-[IVlms]	327–341
4	10	26.1	[IVlgfy]-[SAtmnlfhq]-[Fyla]-[Tvil]-**G**-[Sgen]-[Tsvrindepaqk]-[EAprqgktvnldh]-[VTiasgm]-[Gafi]	364–373
5	16	39.8	[Lamfgs]-[Enlqf]-[Ltmcagi]-[**G**s]-[Ga]-[Knlmqshiv]-[SNadc]-[Pahftswv]-[cnlfmgivahst]-[Ivlyfa]-[Viamt]-[Fdlmhcanyv]-[Daeskprnt]-[Dsntaev]-[Acvistey]-[Dnlera]	397–416
6	8	22.6	[Fyvlma]-[Fgylrmdaqetwsvikp]-[Nhstyfaci]-[QAsnhtcmg]-**G**-[Qe]-[crvitksand]-[**C**r]	430–437
7	9	21.4	[Gdtskac]-[Yfnarthclswv]-[FYlwvis]-[IVlfym]-[Qeapkgrmynhlswyv]-[Pa]-[Tachlmy]-[VIl]-[FLivwn]	533–542
8	7	22.9	[Ektdrqgs]-**E**-[Ivtlnfsp]-**F**-[Ga]-[**P**s]-[Vilcf]	560–566
9	15	33.5	[Nrst]-[Dnaseqtkrcgi]-[TSrvnalcqgik]-[Epdtgqikvrfshyncl]-[Yfkqvm]-[Gpa]-[Lnmv]-[Astgvqcf]-[Agsltfc]-[AGysct]-[VIlfams]-[Fhwyivlem]-[TSag]-[KRnsqteahdp]-[DNsileakt]	586–600
10	12	30.6	[Pasw]-[Fwyahv]-[Gtqs]-**G**-[Fvyesnimtawrq]-[Kgrn]-[mqarelnskghdpt]-[Stm]-[Gfls]-[Ifntlmygshrvq]-[Gdnhrsy]-[Rdpsagkte]	641–654

[a] Motifs are numbered consecutively in order of appearance in the ALDH sequences.

[b] Motifs are given as ProSite patterns. Capitalized letters represent residues that are predominant at each bracketed position. Residues highlighted in bold are conserved in at least 95% of known ALDHs.

sequence has recently been revised to Asn at this position (P.V. Bhat, pers. comm.).

In addition to the 16 residues noted above, 37 residues, including all of the remaining invariant residues identified previously (Hempel et al., 1993), were conserved in at least 80% of the sequences. Thus, only about 10% (53 out of ~500) of all residues in ALDHs are conserved above the 80% level.

Conserved motifs

Twelve of the 16 residues with at least 95% conservation cluster into 7 of the 10 most conserved motifs in ALDHs (Table 1). Motifs were identified statistically using the MEME program (Bailey & Elkan, 1994). This program examines the sequences independently of any previous alignment and looks for segments of identical length within each sequence that show a high degree of similarity and, hence, "information." Thus, the identification of the motifs presented here is independent of the sequence alignment.

Motif 4 covers the essential NAD-binding turn of the Rossmann fold, between β-4 and α-D in the class 3 ALDH structure. The first glycine in this turn, Gly187 [368], is invariant in ALDHs, as well as in the Rossmann folds of several other dehydrogenase families (Lesk, 1995). The only motif with multiple invariant residues, Glu333 [561] and Phe335 [563], is Motif 8. Motif 5 contains both Glu209 [398], conserved in just less than 95% of the 145 ALDHs and proposed to act as a general base (Abriola et al., 1987; Wang & Weiner, 1995), and also the Gly–Gly [400–401] dipeptide boundary between the coenzyme-binding and catalytic domains. Motif 6 includes the invariant Gly240 [434] and the catalytic thiol, Cys243 [437]. The residue nearest to the catalytic thiol, Asn114 [281] (discussed above), is centered in Motif 1, the most conserved motif in ALDHs. The intriguing "U-turn" spanning β-12 and α-14 is encoded in Motif 10 (Liu et al., 1997). A view of the location of each motif in the class 3 structure will be available at our ALDH web site.

Overall, the 10 motifs reside at or near the active site of the molecule (Fig. 2). A large portion of the β-sheet structure is highly conserved vs. very little helical structure. Nearly all motifs contain a turn or loop with a well-conserved small amino acid residue such as glycine, proline, aspartic acid, or asparagine. The well-conserved large hydrophobic amino acid side chains in these motifs often point away from the rest of the motif and appear to anchor these elements to the core of the protein.

Family relationships

A phylogenetic tree (Fig. 3) was generated from the aligned sequences. The tree consists of two main trunks, the "Class 3" and "Class 1/2" trunks, and at least 13 ALDH families (Table 2). ALDH families represent groups of orthologous sequences, which are clearly paralogous to other ALDH sequences. The root was placed at the midpoint of the tree. While there is no reason to believe this root corresponds to an "ultimate ancestral" ALDH, it does mark a fundamental, previously recognized, division in the ALDH family. The existence of these two main ALDH trunks is supported by a recent alignment of 53 ALDHs, with class 1 and 2 ALDHs in a separate branch of the tree from class 3 ALDHs, MMSALDHs, and SSALDHs. Class 1 and 2 ALDHs also have a

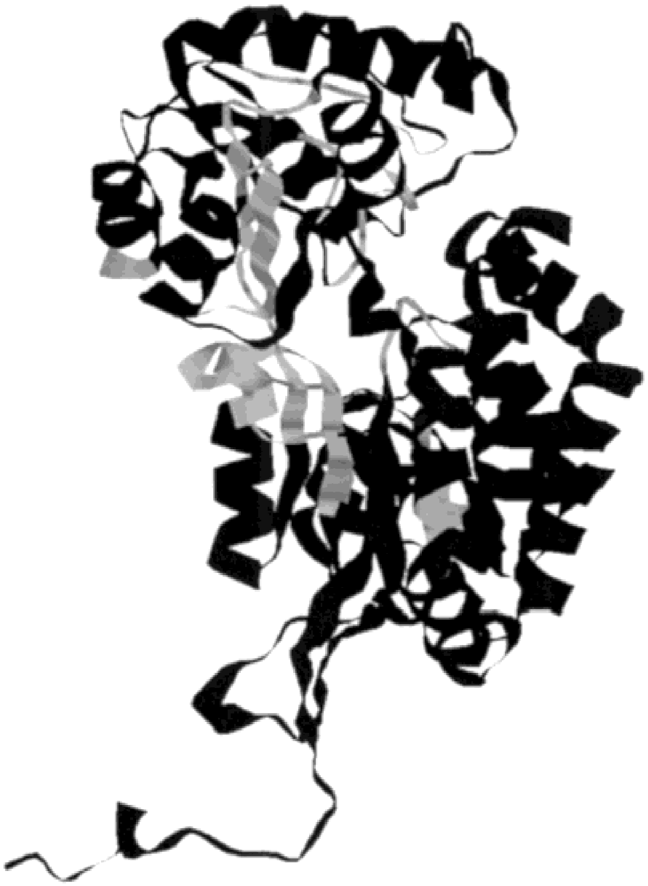

Fig. 2. Conserved motifs in ALDHs. Nine of the 10 motifs presented as ProSite patterns in Table 1 are highlighted in gray in the rat cytosolic class 3 ALDH (black), generated by RasMol (Sayle & Milner-White, 1995). Class 3 ALDHs lack the conservations of Motif 3. Note the clustering of the motifs near the active site, the region between the two domains.

different exon organization when compared to cytosolic and microsomal class 3 ALDHs (Yoshida et al., 1998).

Sequences were assigned to families based on positions within the tree, as well as pairwise sequence identities and evolutionary distances. Evolutionary distance is measured in accepted point mutations per 100 amino acids (PAMs) and is inversely related to pairwise percent identity. As the percent identity increases, the evolutionary distance between the two sequences decreases (Dayhoff et al., 1978). Family assignment was complicated by the lack of data on enzymatic activity for many sequences, most of which were generated by genome sequencing projects. Kolmogorov–Smirnov analysis (Sokal & Rohlf, 1981) of sequences between and within the various ALDH families yields a Dstat value of 0.65, indicating that sequences within each ALDH family are clearly more related to each other than to sequences assigned to the other ALDH families (see the Web site for more details). Currently, we cannot reliably assign some ALDHs to a specific family, but as more sequences are reported and better knowledge of substrate specificity is available, other ALDH families may emerge.

Although far from being an absolute distinction, most families in the "Class 3" trunk are substrate-specific ALDHs, while in the "Class 1/2" trunk, ALDH families with variable substrate specificity are more often found. It is evident that differentiation of a primordial ALDH into the various substrate-specific ALDH families occurred early (Fig. 3). Sequences within these families have since evolved independently from those in other ALDH families. As further indication that substrate-specific types diverged early, some families, such as GGSALDHs and MMSALDHs, include sequences from organisms ranging from bacteria to mammals. Variable substrate ALDHs, including class 1, 2, and 3 ALDHs and the Fungal ALDHs, appear to have diverged much later in evolution. These ALDHs appear to adapt readily to new environments and evolutionary niches. Of these four variable-substrate families, only the class 3 ALDHs appear to have orthologs in a wide phylogeny of organisms, including cyanobacteria, while class 2 ALDHs currently include sequences from only plants and animals and class 1 only animal sequences.

The separation of ALDHs into two trunks cannot adequately explain the diversity of quaternary structures and coenzyme preferences among ALDHs (Table 2). The "Class 1/2" trunk contains ALDH families that are mostly homotetramers, though HMSALDHs are homodimers. Similarly, most ALDHs in the "Class 1/2" trunk utilize only NAD, though FTDHs use NADP. In the "Class 3" trunk, most ALDHs are also homotetramers, except SSALDHs, GGSALDHs, and class 3 ALDHs themselves, and most require NAD as a cofactor, except GAPDHs and a group of bacterial SSALDHs, which use NADP, and class 3 ALDHs, which can utilize either NAD or NADP. Subcellular localization throughout the two trunks is also varied.

New families

Three ALDH families presented here are newly or very recently recognized. Fungal ALDHs consist of a number of variable substrate ALDHs from fungi that are closely related to class 1 and 2 ALDHs. This family includes both cytoplasmic and mitochondrial ALDHs. Based on the tree (Fig. 3B), it is likely that the ancestor of the "Class 1/2" trunk diverged along two separate paths, one leading to the Fungal ALDHs and the other to the higher plant and animal forms, represented by the class 1 and 2 isozymes. As Fungal ALDHs include both cytosolic and mitochondrial forms and as cytosolic and mitochondrial forms are in separate families in animals, it is possible that a cytosolic-mitochondrial schism may have occurred twice during ALDH evolution. If it had occurred only once, it might be expected that the cytosolic Fungal ALDHs would be more closely related to class 1 ALDHs, while the mitochondrial Fungal ALDHs would be more related to class 2 ALDHs.

Another family of ALDHs, which has only recently been noted (Priefert et al., 1997), contains ALDHs that oxidize relatively specific aromatic xenobiotics found in the environment. Some examples of the substrates metabolized by these enzymes are vanillin (Priefert et al., 1997), salicylaldehyde (Denome et al., 1993), and benzaldehyde (Inoue et al., 1995). Benzaldehyde dehydrogenase (Psepu-benz, Acin-benz) is involved in the upper portion of a pathway that also includes another ALDH, 2-hydroxymuconic semialdehyde dehydrogenase (Inoue et al., 1995). Even though these Aromatic ALDHs participate in different catabolic pathways for aromatic compounds, their sequences are closely related to each other, suggesting they may have a recent common ancestor.

Turgor-responsive ALDHs in plants can be induced by osmotic stress (Guerrero et al., 1990; Stroeher et al., 1995). A human protein, antiquitin (Human-ant), is closely related to these plant proteins (~59% identity, 58 PAMs, to each), as well as two hypothetical sequences from *Caenorhabditis elegans* (Celeg-YLQ6) and *Myco-*

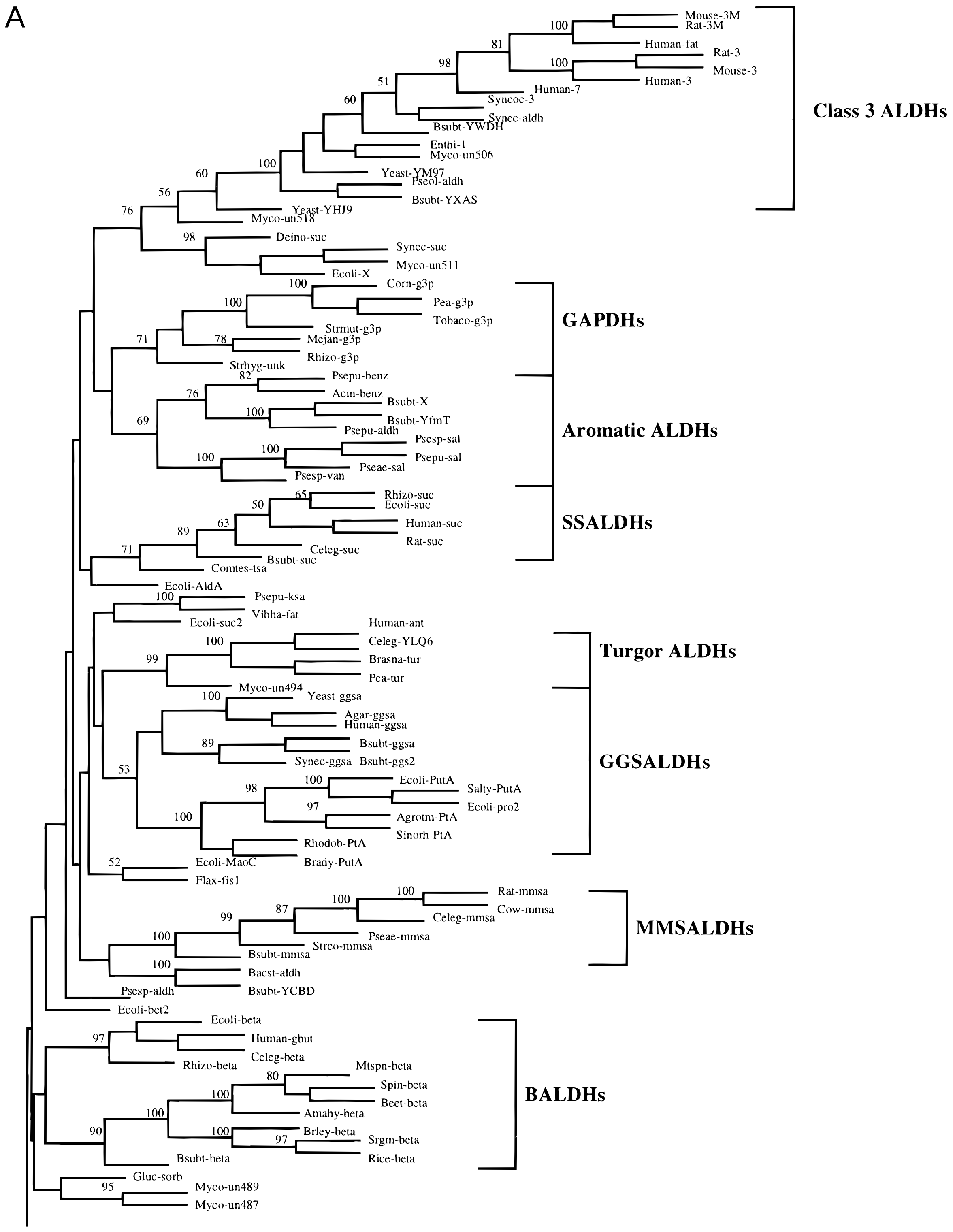

Fig. 3. Phylogenetic tree of known ALDHs. Sequence names and references are available at the ALDH Web site. Sub-branches containing individual ALDH families are bracketed to the right of the branch. The two main branches are (**A**) the "Class 3" branch and (**B**) the "Class 1/2" branch. Bootstrap values are provided to the left of selected branch points to illustrate the close relationship between the ALDHs within each family. (*Figure continues on facing page.*)

bacterium tuberculosis (Myco-un494). Unfortunately, no information on enzymatic activity has yet been reported for any of these proteins. They all possess a rare insertion from indices 328–331 and all but one have a short insertion (indices 601–602), which is also present in GGSALDHs from higher organisms. As Turgor ALDHs are present in the "Amino Acid Intermediate" sub-branch

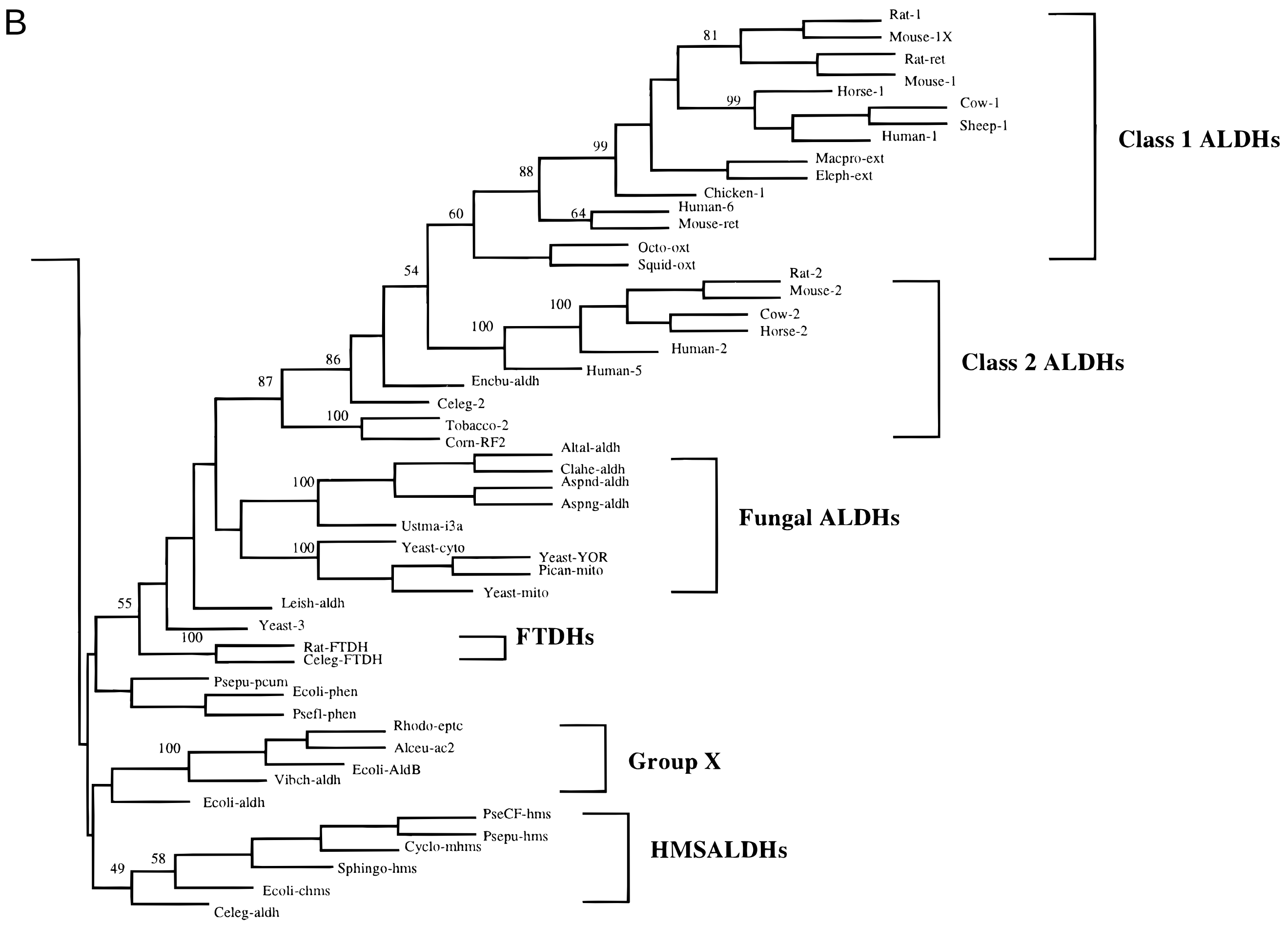

Fig. 3. *Continued.*

of the “Class 3” trunk, it could be proposed that Turgor ALDHs also function in a similar metabolic pathway.

An additional cluster of highly related sequences, designated “Group X” in Figure 3B, is present in the Class 1/2 Branch. These sequences share 61–75% identity (54–30.2 PAMs) to each other with a bootstrap value of 100 for this branch. However, their relation to each other is not yet understood. All four of these enzymes appear to utilize NAD as a cofactor. All appear to oxidize aliphatic aldehydes. Ecoli-AldB has been identified as a lactaldehyde dehydrogenase (Sofia et al., 1994; Xu & Johnson, 1995), while Vibch-aldh and Alceu-ac2 are acetaldehyde dehydrogenases (Parsot & Mekalanos, 1991; Priefert et al., 1992). These ALDHs thus appear to prefer short-chain aliphatic aldehydes. However, Rhdoc-eptc is more active with long-chain aliphatic aldehydes (Nagy et al., 1995). The transcription of the Vibch-aldh and Rhdoc-eptc genes are also induced by toxic compounds. Though Group X likely represents a 14th ALDH family, insufficient functional data exist for these proteins to currently categorize this family. As these ALDHs prefer aliphatic aldehydes and NAD, it may be that these sequences represent class 1/2-like bacterial ALDHs.

Possible subfamilies

Three ALDH families appear to contain separate subfamilies. First, Fungal ALDHs appear to separate into two groups: mold ALDHs (Altal-aldh, Clahe-aldh, Aspnd-aldh, Aspng-aldh) and yeast ALDHs (both cytosolic and mitochondrial).

Next, two groups of betaine ALDHs (BALDHs) are apparent (Fig. 3A). One includes all plant BALDHs, with 62–89% identity (52–11.8 PAMs evolutionary distance) to each other. The other group contains BALDHs from bacteria and *C. elegans* and also human γ-aminobutyraldehyde dehydrogenase (Human-gabt) with 35–53% identity (133–72 PAMs) to each other. Pairwise identities between sequences from the two groups range from 33–42% (143–104 PAMs). Human γ-aminobutyraldehyde dehydrogenase has a low K_m (5 μM) for γ-aminobutyraldehyde and has been postulated to act in the conversion of putrescine to the inhibitory neurotransmitter γ-aminobutyric acid (GABA). However, Human-gabt can effectively oxidize betaine aldehyde (Pietruszko et al., 1997). Based on sequence comparison, Human-gabt is a member of the BALDH family. This enzyme is sometimes referred to as the human E3 isozyme, but is not a class 3 ALDH.

Finally, in most organisms the oxidation of γ-glutamyl semialdehyde is catalyzed by free (which we refer to as type I) γ-glutamyl semialdehyde dehydrogenases (GGSALDHs). Type II GGSALDHs (PutA proteins), encoded by the *putA* gene and identified to date only in bacteria, are actually multifunctional fusion proteins of proline dehydrogenase and full-length GGSALDH. They can convert proline directly to glutamate via γ-glutamyl semialdehyde (Ling et al., 1994).

Table 2. *The 13 ALDH families*

Family	Abbreviation	Substrate specificity	Quaternary structure[a]	Coenzyme(s)
"Class 3" trunk				
Betaine ALDH	BALDH	Specific	Tetramer	NAD
"Amino acid intermediate" sub-branch				
Methylmalonyl semialdehyde DH	MMSALDH	Specific	Tetramer	NAD & CoA
γ-Glutamyl semialdehyde DH	GGSALDH	Specific	Dimer	NAD
Turgor-responsive ALDH	Turgor ALDH	ND[b]	ND[b]	ND[b]
"Class 3" sub-branch				
Class 3 ALDH		Variable	Dimer	NAD or NADP
Aromatic-metabolizing ALDH	Aromatic ALDH	Specific	Tetramer	NAD
Succinic semialdehyde DH	SSALDH	Specific	Tetramer	NAD (animals) NADP (bacteria)
Glyceraldehyde-3-phosphate DH (nonphosphorylating)	GAPDH	Specific	Tetramer	NADP
"Class 1/2" trunk				
Class 1 ALDH		Variable	Tetramer	NAD
Class 2 ALDH		Variable	Tetramer	NAD
Fungal ALDH		Variable	Tetramer	NAD
10-Formyltetrahydrofolate DH	FTDH	Specific	Tetramer	NADP
2-Hydroxymuconic semialdehyde DH	HMSALDH	Specific	Dimer	NAD

[a] All known ALDHs have identical subunits.
[b] Not determined.

The ALDH domain of the PutA proteins is most related to the two type I GGSALDHs from *B. subtilis*, the only bacterial GGSALDH sequences currently available. A sequence from *Synechocystis* (Synec-ggsa), identified as a type I GGSALDH by genomic sequencing, has an extended amino-terminus with homology to the type II GGSALDHs. However, its reported amino-terminus is not as long as the PutA proteins and it lacks a carboxy-terminal extension. Thus, the Synec-ggsa sequence likely represents a portion of a larger PutA fusion protein.

Another ALDH family may consist of two separate branches in the phylogenetic tree (Fig. 3). The relationships of succinic semialdehyde dehydrogenases (SSALDHs) are complex. Eukaryotic SSALDHs use NAD as a cofactor, while bacteria possess two forms of SSALDH, defined by a dependence on either NAD or NADP (Chambliss et al., 1995). The animal NAD-dependent and bacterial NADP-dependent forms group together on the tree (Fig. 3A). However, two other sequences from *Synechocystis* (Synec-suc) and *Deinococcus* (Deino-suc) were reported to be SSALDHs but do not group with the other SSALDHs. These two are closely related to each other and are more closely related to class 3 ALDHs. No enzymatic data for these two proteins have been reported.

ALDH fusion proteins

ALDHs exist not only as free enzymes, but some large fusion proteins have an ALDH domain. As noted above, type II GGSALDHs (PutA proteins) have a GGSALDH domain fused to a proline dehydrogenase domain. In addition, formyltetrahydrofolate dehydrogenase (FTDH) is a large fusion protein of about 900 residues with three domains. Its carboxy-terminal, ALDH domain is closely related to class 1 and 2 ALDHs (Fig. 3B). However, FTDH uses NADP at physiological concentrations while class 1 and 2 ALDHs use only NAD (Cook et al., 1991; Krupenko et al., 1997).

Unique sequence regions

Some ALDH families have unique sequence elements relative to other ALDHs. For example, combined with the class 3 and 2 structures (Liu et al., 1997; Steinmetz et al., 1997), the alignment suggests that plant GAPDHs have a shortened turn between β-4 and α-D, which functions prominently in binding coenzyme.

Methylmalonyl semialdehyde dehydrogenases are the only known ALDHs with a requirement for CoA (Kedishvili et al., 1992). One characteristic sequence pattern in MMSALDHs is between indices 397–404 (Fig. 1) in Motif 5 (Table 1): a replacement of the representative L$\underline{\text{E}}$LGGKSP with x$\underline{\text{x}}$M**GAKNH** (bold residues conserved in all known MMSALDHs). MMSALDH is the only ALDH family that lacks Glu209 [398] (underlined position), the proposed general base (Wang & Weiner, 1995). Formation of a CoA-ester by MMSALDHs may not require this conservation. It has thus been suggested that this residue may be involved in the release of the free carboxylate product (Hempel et al., 1993). MMSALDHs also have the longest insertion in the "U-turn" region from indices 657–658 and lack Gly405 [643] and Gly411 [650]. These glycines may be necessary to stabilize the shorter U-turns of most ALDHs.

Curiously, type I GGSALDH sequences from yeast, mushroom, and human have a Phe at index 373, which is otherwise almost always Gly or Ala in other ALDHs (Fig. 1). This residue lies at the beginning of helix α-D, immediately following the critical coenzyme-binding turn (Liu et al., 1997). The *B. subtilis* GGSALDHs have a Gly at this position, suggesting that this residue change to Phe may characterize GGSALDHs of higher organisms. Type I & II GGSALDHs also have an unusual insertion (6–8 residues), which probably extends the loop between α-D and β-5 (indices 385–391). Therefore, GGSALDHs are likely to have a slightly altered topology near the coenzyme-binding site, relative to most other ALDHs.

Class 3 ALDH family members lack Motif 3 (Table 1) and have rather rare indels at indices 356–357 and 506–510. The extent of the class 3 ALDH family in the phylogenetic tree (Fig. 3A) was defined based on the node prior to the evolution of Pseol-aldh and Bsubt-YXAS. This node has a bootstrap value of 100, indicating that the sequences following this node are always found grouped together in the tree.

With the residue and motif conservation now indicated for the entire ALDH extended family, further investigation will concentrate on conservations within each individual ALDH family. These studies may reveal which residues contribute to the different substrate and coenzyme specificities of these diverse ALDH families.

Materials and methods

ALDH-related sequences were identified using BLAST searches (Altschul et al., 1997) of a "non-redundant" protein database consisting of sequences from the Swissprot, PIR, Genbank, EMBL, and DDBJ databases using the human class 1 (Hempel et al., 1984) and rat class 3 (Hempel et al., 1989) ALDH sequences. Profile searches (Gribskov et al., 1990) based on the previously published alignment of 16 ALDHs (Hempel et al., 1993) were also used to identify related sequences. Only full-length sequences (addressed below) were used when generating the alignment; several apparent partial sequences were omitted. Small preliminary alignments (5–8 sequences) were first made using CLUSTAL W (Thompson et al., 1994) on the VMS front ends at the Pittsburgh Supercomputing Center. When necessary, sections of these alignments were then statistically optimized using SAGA (Notredame & Higgins, 1996). Sequences were combined and aligned manually using GENEDOC (available at www.cris.com/∼Ketchup/genedoc.shtml). The previous alignment of 16 ALDH sequences was used as a starting template since extensive efforts to improve it through use of a variety of computer programs have failed (Leksana, 1995). A consensus sequence was made for each ALDH family using the PRETTY program (Wisconsin Package Version 9.0, Genetics Computer Group (GCG), Madison, Wisconsin). These consensus sequences were aligned for display using GENEDOC. "Index numbers" identify positions in the alignment and are used in the text within brackets to refer to specific locations in the sequences.

The GENEDOC program was also used to generate pairwise identity values and perform Kolmogorov–Smirnov analysis (Sokal & Rohlf, 1981), using sequence regions from index numbers 98 to 672. Multiple data sets for sequence regions between these indices were generated using the SEQBOOT program for bootstrapping analysis. Distances for the datasets were determined by the PROTDIST program using a Dayhoff PAM 250 scoring matrix (Dayhoff et al., 1983). Phylogenetic trees for these datasets were generated by the NEIGHBOR program. A consensus tree was produced using the CONSENSE and DRAWGRAM programs. All these programs are in the PHYLIP suite of programs (Felsenstein, 1990). Bootstrap values indicate the frequency that sequences common to a particular branch point are found grouped together after 100 randomized alignments; a value of 100 indicates the sequences following that node were found to group together 100% of the time. Sequence and family relationships were determined from the pairwise sequence identities and the position of the sequences within the phylogenetic tree. Analysis of conserved motifs was facilitated by the MEME program (Bailey & Elkan, 1994).

Note added in proof

Differences in placement of gap boundaries may be seen in comparison with structure-based alignments. A recent structural superposition published with the cod betaine ALDH structure (Johansson et al., 1998) indicates that a single position change in the alignment of the class 3 ALDHs relative to the rest of the sequences in the motif 3 region would be appropriate. Motif 3 is well conserved outside of the class 3 ALDHs but is systematically different within the class 3 ALDHs. In such cases we will make appropriate changes in the alignment published on the web, which we plan to keep current with new sequences and any other experimental or computational results that relate to the alignment. All such updates will be annotated with the pertinent citations.

Acknowledgments

Protein sequence and evolutionary studies were done through the resources of the Pittsburgh Supercomputing Center and the NIH Division of Research Resources (grant RR06009). Additional support was provided by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) grant AA06985. Dr. P.V. Bhat graciously provided unpublished data.

References

Abriola DP, Fields R, Stein S, MacKerell AD, Pietruszko R. 1987. Active site of human liver aldehyde dehydrogenase. *Biochemistry 26*:5679–5684.

Achatz G, Oberkofler H, Lechenauer E, Simon B, Unger A, Kandler D, Ebner C, Prillinger H, Kraft D, Breitenbach M. 1995. Molecular cloning of major and minor allergens of *Alternaria alternata* and *Cladosporium herbarum. Mol Immunol 32*:213–227.

Altschul SF, Madden TL, Schäffer AA, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res 25*:3389–3402.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *ISMB 2*:28–36.

Brändén C, Tooze J. 1991. *Introduction to protein structure*. New York: Garland.

Brocchieri L, Karlin S. 1998. A symmetric-iterated multiple alignment of protein sequences. *J Mol Biol 276*:249–264.

Chambliss KL, Caudle DL, Hinson DD, Moomaw CR, Slaughter CA, Jacobs C, Gibson KM. 1995. Molecular cloning of the mature NAD(+)-dependent succinic semialdehyde dehydrogenase from rat and human cDNA isolation, evolutionary homology and tissue expression. *J Biol Chem 270*:461–467.

Cook RJ, Lloyd RS, Wagner C. 1991. Isolation and characterization of cDNA clones for rat liver 10-formyltetrahydrofolate dehydrogenase. *J Biol Chem 266*:4965–4973.

Creighton TE. 1993. *Proteins: Structures and molecular properties*. New York: WH Freeman.

Dayhoff MO, Barker WC, Hunt LT. 1983. Establishing homologies in protein sequences. *Methods Enzymol 91*:524–545.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, ed. *Atlas of protein science and structure*. Silver Spring, MD: National Biomedical Research Foundation. pp 345–352, 375.

Denome SA, Stanley DC, Olson ES, Young KD. 1993. Metabolism of dibenzothiophene and naphthalene in *Pseudomonas* strains: Complete DNA sequence of an upper naphthalene catabolic pathway. *J Bacteriol 175*:6890–6901.

Felsenstein J. 1990. *PHYLIP manual*, Version 3.3. University of California, Berkeley: University Herbarium.

Gribskov M, Lüthy R, Eisenberg D. 1990. Profile analysis. *Methods Enzymol 183*:146–159.

Guerrero FD, Jones JT, Mullet JE. 1990. Turgor-responsive gene transcription and RNA levels increase rapidly when pea shoots are wilted. Sequence and expression of three inducible genes. *Plant Mol Biol 15*:11–26.

Habenicht A, Hellman U, Cerff R. 1994. Non-phosphorylating GAPDH of higher plants is a member of the aldehyde dehydrogenase family with no sequence homology to phosphorylating GAPDH. *J Mol Biol 237*:165–171.

Hempel J, Harper K, Lindahl R. 1989. Inducible class 3 aldehyde dehydrogenase from rat hepatocarcinoma and 2,3,7,8-tetrachlorodibenzo-*p*-dioxin-treated liver: Distant relationship to the class and 2 enzymes from mammalian liver cytosol/mitochondrial. *Biochemistry 28*:1160–1167.

Hempel J, Liu Z-J, Perozich J, Rose J, Lindahl R, Wang B-C. 1997. Conserved residues in the aldehyde dehydrogenase family: Locations in the class 3 tertiary structure. *Adv Exp Med Biol 414*:9–13.

Hempel J, Nicholas H, Lindahl R. 1993. Aldehyde dehydrogenases: Widespread structural and functional diversity within a shared framework. *Protein Sci 2*:1890–1900.

Hempel J, Perozich J, Chapman T, Rose J, Liu Z-J, Boesch JS, Wang B-C, Lindahl R. 1999. Aldehyde dehydrogenase catalytic mechanism: A proposal. *Adv Exp Med Biol.* In press.

Hempel J, von Bahr-Lindström H, Jörnvall H. 1984. Aldehyde dehydrogenase from human liver: Primary structure of the cytoplasmic isoenzyme. *Eur J Biochem 141*:21–35.

Hsu LC, Chang W-C, Yoshida A. 1997. Human aldehyde dehydrogenase genes, *ALDH7* and *ALDH8*: Genomic organization and genomic gene structure comparison. *Gene 189*:89–94.

Inoue J, Shaw JP, Rekik M, Harayama S. 1995. Overlapping substrate specificities of benzaldehyde dehydrogenase (the *xylC* gene product) and 2-hydroxymuconic SADH (the *xylG* gene product) encoded by TOL plasmid pWW0 of *Pseudomonas putida*. *J Bacteriol 177*:1196–1201.

Johansson K, El-Ahmad M, Ramaswamy S, Hjelmqvist L, Jörnvall H, Eklund H. 1998. Structure of betaine aldehyde dehydrogenase at 2.1 Å resolution. *Protein Sci 7*:2106–2117.

Jörnvall H. 1977. Differences between alcohol dehydrogenases. *Eur J Biochem 72*:443–452.

Kedishvili NY, Popov KM, Rougraff PM, Zhao Y, Crabb DW, Harris RA. 1992. CoA-dependent methylmalonate semialdehyde dehydrogenase, a unique member of the aldehyde dehydrogenase superfamily. *J Biol Chem 267*:19724–19729.

Krupenko SA, Wagner C, Cook RJ. 1997. Expression, purification, and properties of the aldehyde dehydrogenase homologous carboxyl-terminal domain of rat 10-formyltetrahydrofolate dehydrogenase. *J Biol Chem 272*: 10266–10272.

Leksana A. 1995. Generation and analysis of multiple sequence alignments [Masters Thesis]. University of Pittsburgh.

Lesk AM. 1995. NAD-binding domains of dehydrogenases. *Curr Opin Struct Biol 5*:775–783.

Lindahl R. 1992. Aldehyde dehydrogenases and their role in carcinogenesis. *Crit Rev Biochem Mol Biol 27*:283–335.

Ling M, Allen SW, Wood JM. 1994. Sequence analysis identifies the proline dehydrogenase and delta-1-pyrroline-5-carboxylate dehydrogenase domains of the multifunctional *Escherichia coli* PutA protein. *J Mol Biol 243*:950–956.

Liu Z-J, Sun Y-J, Rose J, Chung Y-J, Hsiao C-D, Chang W-R, Kuo I, Perozich J, Lindahl R, Hempel J, Wang B-C. 1997. The first structure of an aldehyde dehydrogenase reveals novel interactions between NAD and the Rossmann fold. *Nature Struct Biol 4*:317–326.

Nagy I, Schoofs G, Compernolle F, Proost P, Vanderlayden J, de Mot R. 1995. Degradation of the thiocarbamate herbicide EPTC (S-ethyl dipropylcarbamothioate) and biosafening by *Rhodococcus* sp. NI86/21 involve an inducible cytochrome p450 system and aldehyde dehydrogenase. *J Bacteriol 177*:676–687.

Ni L, Sheikh S, Weiner H. 1997. Involvement of glutamate 399 and lysine 192 in the mechanism of human liver mitochondrial aldehyde dehydrogenase. *J Biol Chem 272*:18823–18826.

Notredame C, Higgins DG. 1996. SAGA: Sequence alignment by genetic algorithm. *Nucleic Acids Res 24*:1515–1524.

Parsot C, Mekalanos JJ. 1991. Expression of the *Vibrio cholerae* gene encoding aldehyde dehydrogenase is under control of ToxR, the cholera toxin transcriptional activator. *J Bacteriol 173*:2842–2851.

Persson B, Krook M, Jörnvall H. 1991. Characteristics of short-chain alcohol dehydrogenases and related enzymes. *Eur J Biochem 200*:537–543.

Pietruszko R, Kikonyogo A, Chern M-K, Izaguirre G. 1997. Human aldehyde dehydrogenase E3: Further characterization. *Adv Exp Med Biol 414*:243–252.

Priefert H, Krueger N, Jendrossek D, Schmidt B, Steinbuechel A. 1992. Identification and molecular characterization of the gene coding for acetaldehyde dehydrogenase II (*acoD*) of *Alcaligenes eutrophus*. *J Bacteriol 174*:899–907.

Priefert H, Rabenhorst J, Steinbuchel A. 1997. Molecular characterization of genes of *Pseudomonas* sp. strain HR199 involved in bioconversion of vanillin to protocatechuate. *J Bacteriol 179*:2595–2607.

Sayle A, Milner-White EJ. 1995. RasMol: Biomolecular graphics for all. *Trends Biochem Sci 20*:374–376.

Sofia HJ, Burland V, Daniels DL, Plunkett G, Blattner FR. 1994. Analysis of the *Escherichia coli* genome. V. DNA sequence of the region from 76.0 to 81.5 minutes. *Nucleic Acids Res 22*:2576–2586.

Sokal RR, Rohlf FJ. 1981. *Biometry*. San Francisco, California: WH Freeman & Co. pp 787–794.

Steinmetz CG, Xie P, Weiner H, Hurley TD. 1997. Structure of mitochondrial aldehyde dehydrogenase: the genetic component of ethanol aversion. *Structure 15*:701–711.

Stroeher VL, Boothe JG, Good AG. 1995. Molecular cloning and expression of a turgor-responsive gene in *Brassica napus*. *Plant Mol Biol 27*:541–551.

Thompson JD, Higgins DG, Gibson TJ. 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.

Wang X-P, Weiner H. 1995. Involvement of glutamate 268 in the active site of human liver mitochondrial (class 2) aldehyde dehydrogenase as probed by site-directed mutagenesis. *Biochemistry 34*:237–243.

Xu J, Johnson RC. 1995. Fis activates the RpoS-dependent stationary-phase expression of *proP* in *Escherichia coli*. *J Bacteriol 177*:3166–3175.

Yoshida A, Rzhetsky A, Hsu LC, Chang C. 1998. Human aldehyde dehydrogenase gene family. *Eur J Biochem 251*:549–557.

Zinovieva RD, Tomarev SI, Piatigorsky J. 1993. Aldehyde dehydrogenase-derived omega-crystallins of squid and octopus. Specialization for lens expression. *J Biol Chem 268*:11449–11455.