Protein Science (1993), *2*, 1890–1900. Cambridge University Press. Printed in the USA.

Aldehyde dehydrogenases: Widespread structural and functional diversity within a shared framework

JOHN HEMPEL,[1] HUGH NICHOLAS,[2] AND RONALD LINDAHL[3]

[1] Department of Molecular Genetics and Biochemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15261
[2] Pittsburgh Supercomputing Center, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213
[3] Department of Biochemistry and Molecular Biology, University of South Dakota School of Medicine, Vermillion, South Dakota 57069

(RECEIVED June 17, 1993; ACCEPTED July 28, 1993)

Abstract

Sequences of 16 NAD and/or NADP-linked aldehyde oxidoreductases are aligned, including representative examples of all aldehyde dehydrogenase forms with wide substrate preferences as well as additional types with distinct specificities for certain metabolic aldehyde intermediates, particularly semialdehydes, yielding pairwise identities from 15 to 83%. Eleven of 23 invariant residues are glycine and three are proline, indicating evolutionary restraint against alteration of peptide chain-bending points. Additionally, another 66 positions show high conservation of residue type, mostly hydrophobic residues. Ten of these occur in predicted β-strands, suggesting important interior-packing interactions.

A single invariant cysteine residue is found, further supporting its catalytic role. A previously identified essential glutamic acid residue is conserved in all but methyl malonyl semialdehyde dehydrogenase, which may relate to formation by that enzyme of a CoA ester as a product rather than a free carboxylate species. Earlier, similarity to a GXGXXG segment expected in the NAD-binding site was noted from alignments with fewer sequences. The same region continues to be indicated, although now only the first glycine residue is strictly conserved and the second (usually threonine) is not present at all, suggesting greater variance in coenzyme-binding interactions.

Keywords: active site; aldehyde dehydrogenase; conserved folding; glycine conservation; NAD-binding domain; protein evolution; protein family

It is widely appreciated that protein sequences reflect evolutionary history. More recently it has been appreciated that although relationships eventually become blurred at the sequence level, chain-folding patterns are more resistant to change. Some relationships have been detected only after tertiary structural determinations; an early example of this concept is found in the "Rossmann fold" of NAD-binding dehydrogenases (Rossmann et al., 1974), in which sequence similarities were detected more in retrospect (Wierenga & Hol, 1983). Between those relationships seen with ease at the sequence level and those detected only through similar folding patterns are the large families of proteins that usually have broadly similar properties but that share only few strictly conserved residues. Problematic pairwise alignment of the most distantly related members from such families is facilitated by successive alignments starting with less divergent family members (cf. "short chain" alcohol dehydrogenases [Persson et al., 1991]). The present study examines the metabolically related aldehyde dehydrogenases in the absence of any tertiary structure, in order to assess conservations suggesting function.

Aldehyde dehydrogenases (AlDH, EC 1.2.1.3) occur with wide phylogenetic distribution. In mammals, distinct structural classes are noted with specific subcellular distributions but generally without strict organ specificity. They are NAD-linked enzymes (with some NADP-accepting examples) that act on a broad variety of aldehyde substrates, converting them to the corresponding carboxylic acid. In humans, the enzyme is most widely known for its involvement in conversion of ethanol-derived acetaldehyde to acetate and for conversion of biogenic amine-derived aldehydes to their corresponding carboxylic acids. By 1988, primary structures were known for the human and horse mitochondrial and cytosolic AlDH pairs, as well as the rat dioxin- (or TCDD-) inducible AlDH (von Bahr-Lindström

Reprint requests to: John Hempel, Department of Molecular Genetics and Biochemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15261.

et al., 1984; Hempel et al., 1985, 1989; Johansson et al., 1988). Since 1988, the primary structures of a number of additional AlDHs from distant species have been reported. The spectrum now ranges from bacterial forms through examples from fungi and higher plants. In addition to "classical" AlDHs (i.e., enzymes that oxidize a wide variety of aldehyde substrates), structures of enzymes with restricted aldehyde substrate preferences have been added to the family, and new members have been proposed (Kurys et al., 1992). Also, sequences of proteins associated with other functions have been shown to be aldehyde dehydrogenases (e.g., the bovine corneal protein BCP54 [Cooper & Baptist, 1991], androgen-binding protein [Pereira et al., 1991], and a retinal positional marker protein [McCaffery et al., 1991]).

Various pairwise alignments of AlDHs have revealed 32–95% amino acid identities, yet a detailed comparison of all reported structures has been lacking. An earlier account of aligning several AlDHs was reported by Lindahl and Hempel (1991). Since that time, the number of available sequences has doubled, filling phylogenetic gaps and diminishing the number of strict consensus residues. The present group is also more diverse, including enzymes with more narrow substrate preferences (especially semialdehydes), viz. succinate semialdehyde dehydrogenase (from *Escherichia coli*; for references, see Methods), methylmalonyl semialdehyde dehydrogenase (rat), formyltetrahydrofolate dehydrogenase (rat), hydroxymuconic semialdehyde dehydrogenase (*Pseudomonas*), and γ-glutamyl semialdehyde dehydrogenase (yeast).[1] Also, the present analysis avoids presentation of species variants, with at most three exceptions. The phenobarbital-inducible rat enzyme and the human cytosolic enzyme are closely related and both are "class 1" enzymes (in the nomenclature of mammalian AlDHs with broad substrate specificity) with 83% identity. The human AlDHx sequence is clearly a "class 2" species like the liver mitochondrial form, and the rat liver microsomal enzyme is a "class 3" structure, distinguished from its soluble counterpart primarily by a hydrophobic, 17-residue, presumed membrane anchor at the C-terminus (Miyauchi et al., 1991).

Results

In length the AlDH sequences aligned range from the 452-residue TCDD-inducible class 3 structure to the 575-residue yeast γ-glutamyl semialdehyde dehydrogenase. Formyltetrhydrofolate dehydrogenase is 902 residues in its entirety; the C-terminal AlDH domain is given as 486 residues. The alignment required 640 positions to accommodate all gaps and extensions at the N- and C-termini. Here, to distinguish the positions in the alignment from actual position numbers of the individual sequences, we refer to *index numbers* in the alignment and also note a few actual position numbers that are cited frequently.

The alignment (Fig. 1) was initiated using the University of Wisconsin GCG Pileup program with default gap penalties (Genetics Computer Group, 1991). Particularly at the N-terminal end, manual adjustments were introduced because uniform gap penalties are seldom adequate across the entirety of a group of sequences. Overall, alignment was based primarily on residue identity and gap minimization without recourse to aids such as conservation of hydrophobicity or minimum codon change (Argos, 1989). Excluding the species variants noted above, we found pairwise identities from 17% (rat class 3 AlDH vs. yeast γ-glutamyl semialdehyde dehydrogenase) to 68% (human class 1 vs. class 2 enzymes) (Table 1). The exact placement of some gaps may be equivocal, particularly near the N-terminal end, e.g., a nearly invariant Ala might be suggested at index 108 through further gapping of the class 3 and *Pseudomonas* sequences. Instead, we have sought to avoid overaligning to produce artificially contrived residue conservations.

The sequences share a relatively common core, although not without gaps, from indexes 101 to 584, and if the class 3 structures and *Pseudomonas* AlDH are excluded, this core begins around index 53, where all other structures display elements of the "consensus" sequence **lfingew** (Fig. 1). The amino-terminus of the class 3 forms corresponds to the beginning of the third exon of the class 1 and 2 mammalian genes (Hsu et al., 1988, 1989). Functionally, the longer N-terminal domains are suggested by limited proteolysis to be required for tetramer formation (Loomes & Jörnvall, 1991). Only the γ-glutamyl semialdehyde dehydrogenase sequence (and precursors of mammalian liver mitochondrial AlDH [Hsu et al., 1988] and AlDHx [Hsu & Chang, 1991] and spinach betaine-AlDH [Weretilnyk & Hanson, 1990], with N-terminal targeting sequences, not shown here) extend further N-terminally than index 39. At the C-terminus, again the *Pseudomonas* and class 3 enzymes provide the major deviations from the core, with extensions. Short portions of these extensions show similarities to the noncoding portion of the class 2 cDNA, possible evidence of stop codon migration (Hempel et al., 1989).

The alignment indicates strong pressure to maintain a common core throughout members of this family because most gaps after index 200 are forced by insertions in just one or two sequences. The longest gap is 10 residues, at index 400. Twenty-three invariant residues are found (Fig. 1, reverse background). More than half of these are Gly (11) or Pro (3), most likely reflecting critical chain-bending points. The remaining invariant residues are two Asn and one each Cys, Lys, Arg, Glu, Thr, Ser, and Phe. An additional 12 and 15 "nearly invariant" residues, at 93 and 87% conservation (15/16 and 14/16 identities), are

[1] The possibility of sequence similarity between glyceraldehyde-3-phosphate dehydrogenase and AlDH was examined once the first AlDH structures became available, without detection of homology (Hempel et al., 1984).

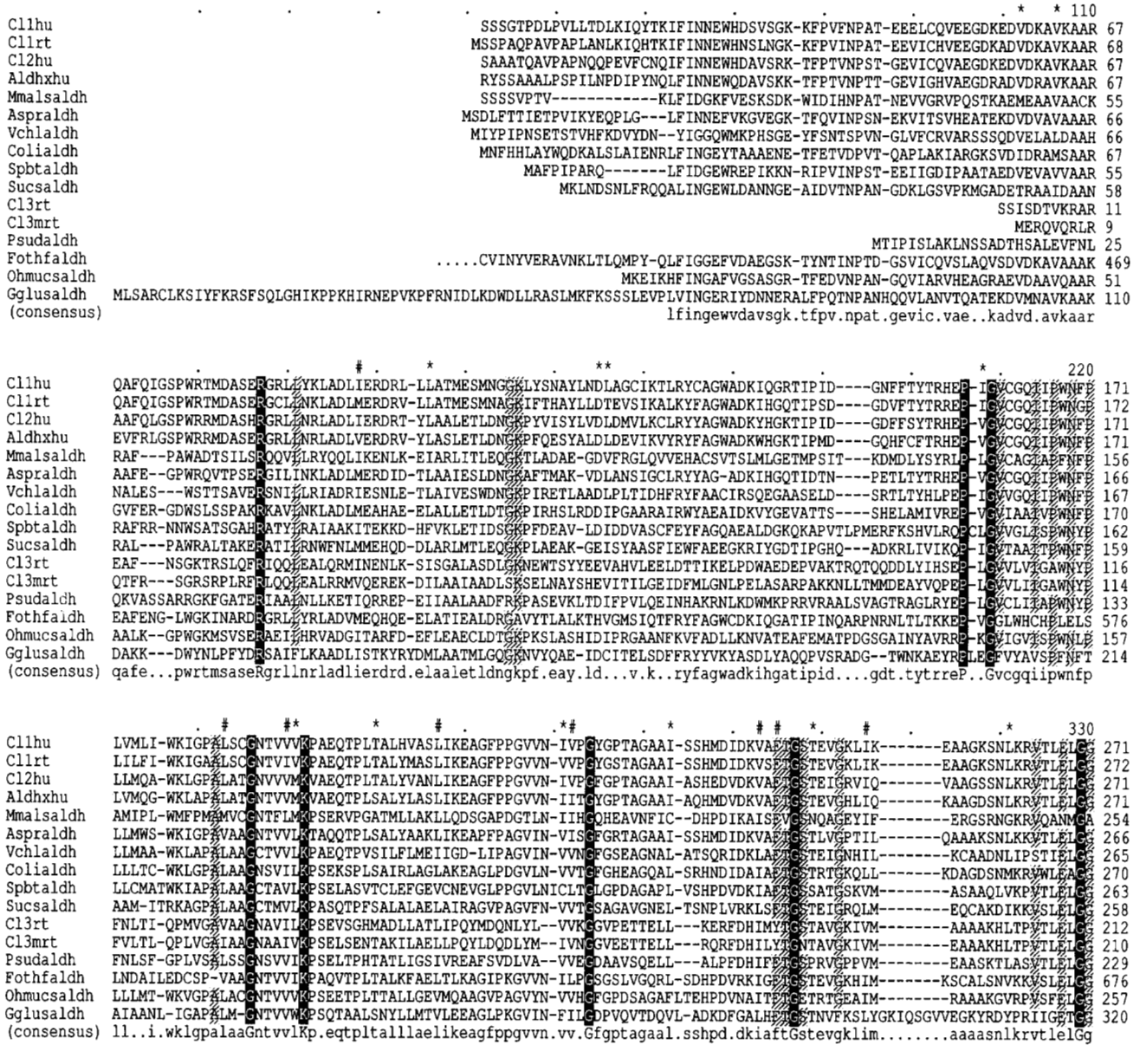

Fig. 1. Multiple alignment of 16 aldehyde dehydrogenase sequences. From the top, the AlDH sequences aligned are: **Cl1hu**, from human liver cytosol (class 1); **Cl1rt**, from phenobarbital induction of rat liver; **Cl2hu**, from human liver mitochondria (class 2); **Aldhxhu**, from human genomic library screening ("AlDHx"); **Mmalsaldh**, methyl malonic semialdehyde dehydrogenase from rat liver; **Aspraldh**, from *Aspergillus*; **Vchlaldh**, from *V. cholera*; **Colialdh**, from *E. coli*; **Spbtaldh**, from spinach (a betaine aldehyde dehydrogenase); **Sucsaldh**, succinic semialdehyde dehydrogenase from *E. coli*; **Cl3rt**, from TCDD induction of rat liver (class 3); **Cl3mrt**, from rat liver microsomes (a membrane-bound class 3 form); **Psudaldh**, from *Pseudomonas*; **Fothfaldh**, the C-terminal half of formyl tetrahydrofolate dehydrogenase, a chimeric protein from rat liver (N-terminal truncation noted by . . .); **Ohmucsaldh**, hydroxymuconic semialdehyde dehydrogenase from *Pseudomonas*; and **Gglusaldh**, γ-glutamyl semialdehyde dehydrogenase from yeast. Strict consensus residues are printed in reverse background, identities at the 93% (15/16) and 87% (14/16) levels are highlighted by hashing. "Invariant similarities," where all residues are K/R, D/E, T/S, Y/F, N/Q, or L/I/V/M are indicated by # above the first sequence. Similarities conserved within one of the above categories of alternatives at the 87% (14/16) level or greater are indicated by *. Index numbers in increments of 10 are indicated by dots on the top line, with the actual index number given at the end of each top line. The actual position number of the last residue of each individual sequence in each row is given in the right margin following that residue. (*Continues on facing page.*)

denoted by gray and hashed highlighting, respectively. Of these 27 positions, five involve Gly and two involve Pro; otherwise, no other residue is counted more than twice in either category and there is a relatively even balance between numbers of polar, nonpolar, and charged residues. Interestingly, because catalytic participation of His and/or Tyr in AlDH was once supported (Takahashi et al., 1981), and Trp is weighted highly in scoring matrices, there are no

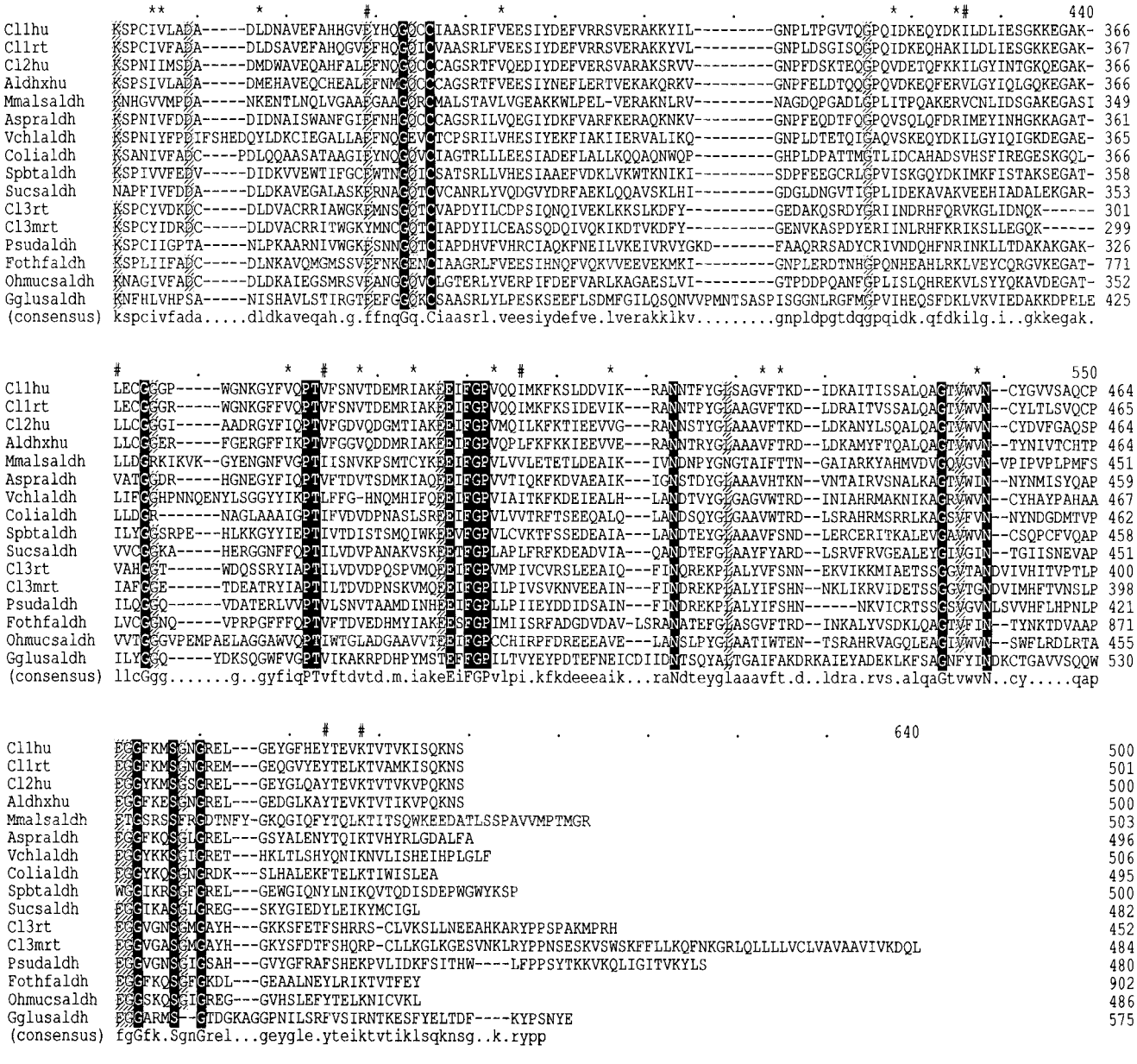

Fig. 1. *Continued.*

His, Tyr, or Trp residues among the strict or near consensus residues.

Fifteen other "invariant similarities" are noted, where all residues are from one of the following groups: L/I/V/M, R/K, F/Y, D/E, and S/T (Fig. 1, #). Eleven of these involve hydrophobic residues L/I/V/M, whereas three are F/Y and one is R/K. A further 26 similarities at the 87% (14/16) level, using the above categories, are also noted by asterisks (Fig. 1). Twenty-one of these are of the L/I/V/M group, three T/S, and one each basic and acidic conservations, suggesting that most of these similarities are important for interior-packing interactions. A total of 89 positions are noted in one of the above ways (indexes 295 and 359 are included in two categories), or about one-seventh of each sequence. No consensus residues are found upstream of index 126. However, one residue in this region is noted: Ile at index 65 is found in all sequences except the class 3 and *Pseudomonas* sequences, which are N-terminally truncated in this region.

Predicted secondary structures

We used the Chou–Fasman parameters to predict secondary structures for these sequences, from indexes 62 to 584, in an effort to determine any correlation between conservations and common, strongly suggested elements of secondary structure. These data are compiled in Figure 2, as average α-helical, β-strand, and reverse-turn potentials.

Table 1. *Percent pairwise identities between aldehyde dehydrogenase sequences*[a]

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Cl1hu	100	82.8	68.2	64.4	30.4	56.5	39.6	38.4	39.4	36.1	26.8	22.7	26.3	45.5	36.6	25.0
2	Cl1rt		100	65.6	63.6	30.5	56.3	38.3	39.4	40.0	36.9	26.6	22.3	26.3	44.7	38.7	24.2
3	Cl2hu			100	73.8	29.6	57.1	40.8	41.0	41.0	35.3	25.9	21.5	26.0	43.7	38.1	24.2
4	Aldhxhu				100	29.6	54.8	41.2	41.8	39.8	35.3	25.7	20.9	25.0	45.5	36.8	24.2
5	Mmalsaldh					100	25.8	21.7	26.7	27.4	27.2	21.7	18.0	21.5	22.2	29.2	23.1
6	Aspraldh						100	42.1	41.0	39.3	36.3	26.1	23.1	24.0	44.6	37.7	26.0
7	Vchlaldh							100	37.6	36.4	32.4	23.5	20.5	21.9	34.7	35.6	21.9
8	Colialdh								100	34.1	35.1	27.4	22.7	22.7	34.6	39.9	26.1
9	Spbtaldh									100	36.3	26.8	23.8	24.8	32.9	36.6	24.6
10	Sucsaldh										100	23.7	21.4	23.5	35.3	36.1	23.9
11	Cl3rt											100	63.3	39.4	22.4	24.6	18.6
12	Cl3mrt												100	38.5	20.3	22.3	17.4
13	Psudaldh													100	21.9	25.0	19.0
14	Fothfaldh														100	34.8	23.7
15	Ohmucsaldh															100	26.3
16	Gglusaldh																100

[a] For abbreviations, cf. Figure 1.

As indicated, an antiparallel helix pair is suggested by helical character from indexes 95 to 115 and 130 to 150. Then, starting at index 210, alternating β and α segments are generally apparent, often punctuated by turns, until index 540, suggesting that AlDHs are predominantly β/α structures.

Discussion

The 16 AlDHs compared here represent a clearly interrelated family, with higher divergence of some pairs but little clear segregation of structures into two or three more closely related groups (Table 1; Fig. 4). Greater conservation at the C-terminal half of the sequences is also seen in Figure 1, where 18 of 23 invariant residues are found after index 320. This greater downstream conservation may reflect greater specific requirements of the catalytic domain because this position marks the major site of limited tryptic cleavage of human AlDHs, separating the apparent nucleotide-binding and catalytic domains (Loomes & Jörnvall, 1991). The catalytic domain may have an extremely ancient relationship with other thiolesterases, as seen from alignment of AlDHs with thiol proteases and correlations with the tertiary structure of papain (Hempel et al., 1991).

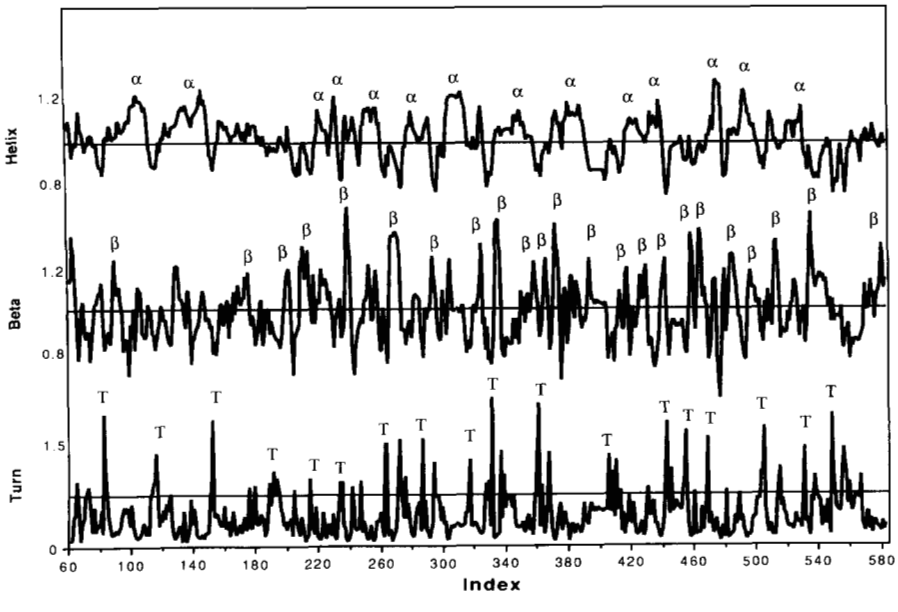

Fig. 2. Average secondary structural potentials using Chou–Fasman parameters, derived from the aligned sequences of Figure 1. Data are plotted as average values (open circles) with standard deviations (bars). Horizontal lines correspond to the threshold values taken to reflect confidence in assignment of respective predictions. Best estimates from these data are summarized in the bottom line as α, β, or T for helix, strand, or turn segments, respectively.

Conservation of tertiary structure in preference to primary structure has become a common theme in protein evolution (Brändén & Tooze, 1991) and is further suggested by the present alignment. It might be expected a priori that the most highly conserved residue clusters of primary structure (for instance, those segments contributing to the consensus sequence **vtlelGgkspciv** at indexes 324–336) would, artificially, provide the strongest suggestion of one type of secondary structure. In the AlDHs, such conserved segments provide no strong suggestion of any regular secondary structure because most of these segments are themselves of low α/β potential. Considering the anti-helical and anti-strand potentials of Gly and Pro, and to a lesser degree Asn, Thr, and Ser (Chou & Fasman, 1978), which together constitute 17/23 invariant residues, this may not be surprising.

Highly conserved segments are expected to play key roles in catalytic function. However, because the assembly of binding sites from residues located on turns or loops is frequent (Fetrow & Rose, 1990), it might in fact be more surprising if the most highly conserved regions were suggested to form regular structure. The α-helical, β-strand, and turn potentials of segments such as **fxnxGqxCia** (indexes 359–368) and **pfgGfkxSgxGr** (indexes 550–561) give the impression of forming specialized random (nonrepetitive) secondary structure, including reverse turns (see below), which would be consistent with this theme. As illustrated by the globin family (Lesk & Chothia, 1980), if maintenance of a simple α-helix or β-strand is crucial, many residues can substitute for all but the most important one(s), thus such segments may not contain any residue conservations. The strong helical averages of segments at indexes 94–114 and 130–148, where there are no invariant residues and only a few positions with a notably high level of conservative replacements, may reflect similar structural requirements.

Earlier recognized functional residues

At index 366, the invariant cysteine residue, Cys 302 of mammalian classes 1 and 2 AlDH, finds continued support in a catalytically essential role. Of the many cysteine residues present in the various individual AlDH structures, only this one is strictly conserved. It is labeled selectively by iodoacetamide and implicated in the reaction of AlDH with disulfiram (Hempel et al., 1982), and is also labeled by a reactive (brominated) coenzyme analog (von Bahr-Lindström et al., 1985) and a vinyl ketone analog of a long-chain (insect pheromone) aldehyde (Pietruszko et al., 1991). As such, this residue is now generally regarded as the catalytic cysteine.

Another important residue was also identified earlier by chemical modification: Glu 268 (index 327) in human liver AlDHs is modified selectively by bromoacetophenone (Abriola et al., 1987) and finds continued support from its high conservation. However, the acidic functionality is lost at this position in methylmalonyl semialdehyde dehydrogenase (Kedishvilli et al., 1992), where an asparagine residue occurs. Interestingly, of all the AlDHs compared, this enzyme is the only one known to produce a CoA ester rather than a free carboxylic acid from a free aldehyde, tempting the speculation that Glu 268 is required for expulsion of a free acid product, but not when the mechanism involves thiolester exchange (attack on the enzyme-bound thiolester intermediate by another thiol), to form product directly.

The alcohol-sensitizing mutation

At index 576, a residue found through natural mutation to be incompatible with catalytic activity is the lysine (in place of glutamic acid) residue at position 487 of human liver mitochondrial AlDH from individuals of Asian ancestry with intolerance to ethanol. This mutation is known to be dominant because heterozygotes also display this sensitivity (Crabb et al., 1989). Their class 2 AlDH activity is nil despite expression of both types of subunit, presumably the result of subunit–subunit interactions, because only 6% of all tetrameric species from equimolar amounts of randomly associating subunits in a heterozygote would contain four nonmutant subunits. The specific purpose of this glutamate moiety has never been suggested. The present alignment, still with flanking similarities at this location, viz. Tyr–Phe at index 574 and Lys–Arg at index 578, contributes little to any explanation because 9 of the 16 sequences contain residues other than Glu at index 576 (Glu, Gln, Asn, His, Arg, and Ser).

A probable coenzyme-binding region

The nucleotide-binding domain of aldehyde dehydrogenase has been localized to the amino-terminal half of the protein by limited proteolysis (Loomes & Jörnvall, 1991). Consistent with this suggestion, segments from indexes 295 to 327 from various individual AlDH sequences have been noted previously to correspond most closely to the residue pattern characteristic of a portion of the nucleotide-binding domain. The expected pattern is GXGXXG (Wierenga & Hol, 1983), taken to reflect the turn at the end of the first β-strand in the first mononucleotide-binding unit of the Rossmann fold, which interacts with the adenine ribose of NAD. With (save once) a threonine in place of the second glycine residue, the sequence **ftGstevg** at indexes 295–302 provides the closest match to this pattern. The residue at index 302 would be an invariant Gly, except for the Phe from γ-glutamyl semialdehyde dehydrogenase. As seen in Figure 1, that sequence is the only one of the 14 structures to have an insertion just after this point, suggesting addition of a loop. Furthermore, although far from fitting exactly the core pattern "expected" of such segments, the segments flanking these glycine residues seem to have many of the expected ele-

k i a f t G s t e v g k x i m x x a a x x n l k - - x v t l - e l

x Ø x Ø x G x G x x G x x x Ø x x Ø x x x x x x x x x Ø x Ø x ⊖ x

Fig. 3. Putative coenzyme-binding domain of aldehyde dehydrogenases aligned against the expected pattern (Wierenga & Hol, 1983). Upper sequence: consensus sequence derived from segments at index 293–329 (Fig. 1) but excluding the gap forced by the γ-glutamyl semialdehyde dehydrogenase sequence. Lower sequence: GXGXXG-containing pattern expected of segments from nucleotide-binding folds. Ø denotes a hydrophobic residue and ⊖ denotes an acidic residue; residues underlined are the same as those found in at least 3 of 11 coenzyme-binding segments compiled recently (Brändén & Tooze, 1991).

ments (Wierenga & Hol, 1983; cf. Fig. 3). Interestingly, the apparent canonical acidic residue at the C-terminal end of the coenzyme-binding pattern is Glu 268, yet the residue at this position is not expected to be acidic in NADP-specific enzymes (Brändén & Tooze, 1991). The two class 3 AlDHs readily use NADP, although they are not obligate in this respect, and each still has a glutamate residue at this position. These observations show that canonical sequences can be quite useful, but that some caution may still be prudent in describing this segment as the site of the crucial turn in the Rossmann fold on the basis of residue patterns alone,[2] supporting the view that "most consensus sequences are not quite essential and almost never suffice to identify a structure–function relationship" (Doolittle, 1989, p. 608). The recent finding that porcine lens aldose reductase is an α/β barrel protein (Rondeau et al., 1992) provides another cautionary note against the expectation that AlDHs even have a Rossmann fold. Based on secondary structural predictions, it is difficult to distinguish between open β/α structures (e.g., with Rossmann folds) vs. closed barrels because both have alternating β and α segments.

Regardless, the implication that the segments at indexes 290–362 form the first mononucleotide-binding unit seems strengthened by consideration of the predicted secondary structures. The segment **ftGs** (indexes 295–298) yields the strong suggestion, with low standard deviation, of a turn. Just upstream, a short β-strand is indicated convincingly, whereas just downstream of the gap forced by γ-glutamyl semialdehyde dehydrogenase, two turns of α-helix are indicated. Continuing further downstream, a turn and alternating β and α segments are suggested, followed by a pair of turns flanking Cys 302 at index 366. These indications are quite compatible with the overall expectation that the catalytic residue should lie in proximity to the coenzyme. Thus, starting from the third glycine residue of the presumptive "GXGXXG" at index 302, an $\alpha\beta\alpha\beta$ segment would carry the main chain away from and back toward this point twice, compatible with location of the coenzyme near the catalytic cysteine rather than on opposite faces of the enzyme.

[2] Another glycine-rich segment is also noted, at indexes 566–573, but this seems less likely as a potential coenzyme-binding site on both primary and predicted secondary structural grounds.

Critical chain-bending points

The abundance of consensus glycine residues in addition to those in the apparent coenzyme-binding site is consistent with the long but often underappreciated importance of glycine in allowing bends outside the limits imposed by side chains of all other residues (Neurath, 1943). This points to a chain-folding pattern changing much more slowly than individual residues, characteristic of many protein families.

A four-residue segment, almost always with two glycine residues (**sgxG** at indexes 557–560), is predicted as a turn, as are many of the other invariant glycine residues in Figure 2. Three of these are found in nearly invariant Gly–Gly pairs (indexes 329–330, 444–445, and 552–553), but only the second one of these pairs is indicated to be the center of a reverse turn, despite the necessity of one and frequently two glycine residues in type I′, II, and II′ (or "inverse common," "glycine," and "inverse glycine") turns (Richardson & Richardson, 1987, 1989). According to Creighton (1993), Gly–Gly is required at positions $i + 1$ and $i + 2$ in the rare type I′ and III′ turns. In liver alcohol dehydrogenase, two Gly–Gly segments participate in severe bends in the peptide chain without being part of typical reverse turns.[3] Thus, these Gly–Gly segments in the AlDH family seem to be required for bends more specialized than reverse turns. In addition, there are several instances of coinciding Gly in all but one or two of the 16 sequences: cf. indexes 302 (in the putative coenzyme-binding fold, above), 330, 415, 445, 552, and 558. Similar observations regarding overrepresentation of consensus glycine residues within the large sets of sequences now

[3] As judged from ϕ, ψ angles of −80°, −165° and −67°, −43° for Gly 201–202, and 60°, 22° and 76°, 0° for Gly 320–321 (Eklund et al., 1976; Bernstein et al., 1977; Abola et al., 1987; from Brookhaven data entry 8ADH, version of April 15, 1991), these Gly–Gly segments permit a main-chain bend far outside the limits allowed with other residues. Using these examples, and using the bond angle specifications for reverse turns, the bond angles of Gly 201 and Gly 202 lie outside even the classified turn types, in a generally forbidden portion of the Ramachandran plot described as permissible only with Gly and then only "with slight flexibility of bond angles" (Creighton, 1993). As residues $i + 1$ and $i + 2$, the bond angles of Gly 320 and Gly 321 fit the specifications of a type I′ turn, yet the dihedral angle formed by the four α-carbon atoms is −71° (vs. −45° for a perfect type I′ turn) and the i to $i + 3$ carbonyl to amide distance is 3.3 Å, vs. the 2.9-Å perfect hydrogen-bonding distance. Neither pair was described originally as forming part of a reverse turn, nor do these values indicate such a structure.

available for both long- and short-chain alcohol dehydrogenases have been made (Jörnvall, 1977; Borrás et al., 1989; Persson et al., 1991).

Another segment with an invariant Gly is **EiFGP** at indexes 478–482, which seems likely to terminate the C-terminal end of an α-helix. This segment also provides the only example of three contiguous invariant residues, yet any suggestion of the specific significance of these residues remains to be detected. The phenylalanine residue, at least, suggests a crucial site for hydrophobic interaction just downstream of a polar environment.

Internal packing

Whereas the most frequently conserved residue was found to be glycine, the large majority of invariant and highly-conserved similarities were the hydrophobic amino acids L/I/V/M. Correlation of the locations of these amino acids, as indicated in Figure 1, with the predicted secondary structures in Figure 2, shows that they frequently occur in locations with high β-strand potential, particularly in short runs of contiguous hydrophobic amino acids (cf. indexes 240, 241, 271, 272, 335, 336, 483, 486, 535, 537). This suggests buried parallel β-strands in $\beta\alpha\beta$-folding units, consistent with the architecture of other known NAD-binding dehydrogenases. All the above positions occur as pairs within a given predicted strand, and all but the 535, 537 pair are both odd and even numbered, suggesting important hydrophobic packing interactions pointing from both faces of a β-sheet.

General conservation of Cys and Trp

In searching protein sequence data banks, conservation of Cys or Trp is usually more greatly rewarded than any other residue because the one of the Dayhoff PAM substitution matrices (Dayhoff et al., 1983) is usually the default option in alignment programs. This matrix is nearly 15 years old and has been recalculated recently based on the now much expanded set of available protein sequences (Jones et al., 1992), with many substitutions involving Cys or Trp being penalized far less severely. The residue conservations seen here are consistent with this reevaluation. Thus, with the exception of Cys 302, no more than seven cysteine residues coincide at any position in the alignment (index 211), whereas the greatest conservation of tryptophan is in 13/16 sequences at indexes 119 and 217. At the first of these locations, Trp is found just downstream of a gap in many of the sequences, suggesting the location of a turn (because gaps are accommodated most easily in turns or loops) even though it is not predicted as such, whereas at the latter location it occurs in a pentapeptide segment (indexes 216–220) usually punctuated by proline residues, and with above-threshold turn potential. Considering the likelihood of tryptophan to occupy the fourth position of a reverse turn (Chou & Fasman, 1978), the function of both of these Trp residues may relate to turn formation.

Evolution

Clearly, the alignment supports divergence of these 16 structures from a common ancestor (or, in the case of formyl tetrahydrofolate dehydrogenase, recruitment of a part of its structure from this common ancestor). An unrooted family tree, generated by the FITCH method in the PHYLIP suite of programs (Fitch & Margoliash, 1967; Felsenstein, 1990) from the "common core" of the alignment, between indexes 102 and 584, is given in Figure 4. The six mammalian AlDH sequences already noted to be class 1, 2, and 3 pairs are expectedly close, and the class 1 and 2 groups are more closely related to each other than any of the other examples. Otherwise, and considering each of these mammalian pairs as one entry each, any pair of sequences appears nearly as divergent as any other, showing that the classification system for mammalian "classical" AlDHs does not extend to more diverse species. Although AlDHs of both broad and narrow substrate specificities are represented here, and they do not segregate accordingly on the tree, the broadly divergent nature of this family of interrelated structures is reinforced.

Beyond reliance on computer programs to determine the apparent separation of these enzymes, the closer relationship of the *Pseudomonas* enzyme to the class 3 AlDHs is noted visually. These three are both truncated at the N-terminal end and extended at the C-terminal end to a greater extent than the other sequences. They also share a pair of gaps, common to just these three sequences, between indexes 269 and 285, as well as the segment E/DKPLALY, at indexes 514–520. No other closer groupings are noted, underscoring the relatively even interrelatedness of the 16 structures compared here, as suggested by the unrooted tree. For this reason we refer to these enzymes as a family and not a superfamily, where distinct families could be identified through clustered relationships and divergent enzymatic functions, such as with the recent suggestion that aldehyde dehydrogenases and thiol proteases may be related through origins from a common thiolesterase (Hempel et al., 1991).

Regarding this putative thiolesterase relationship, residues at 16 positions were noted as conserved or highly similar between four thiol proteases and four AlDHs. Seven of those residues were indicated as strictly conserved, 5 of which are among the 23 invariant residues of Figure 1. One of the folding relationships suggested from that study was maintenance of an internal salt bridge, similar to that between Glu 35 and Lys 174 in papain. That conservation was maintained in all but class 3 AlDH, where residues at both corresponding positions were replaced by uncharged alternatives. Examination of residues at indexes 377 and 556 in the present study reveals

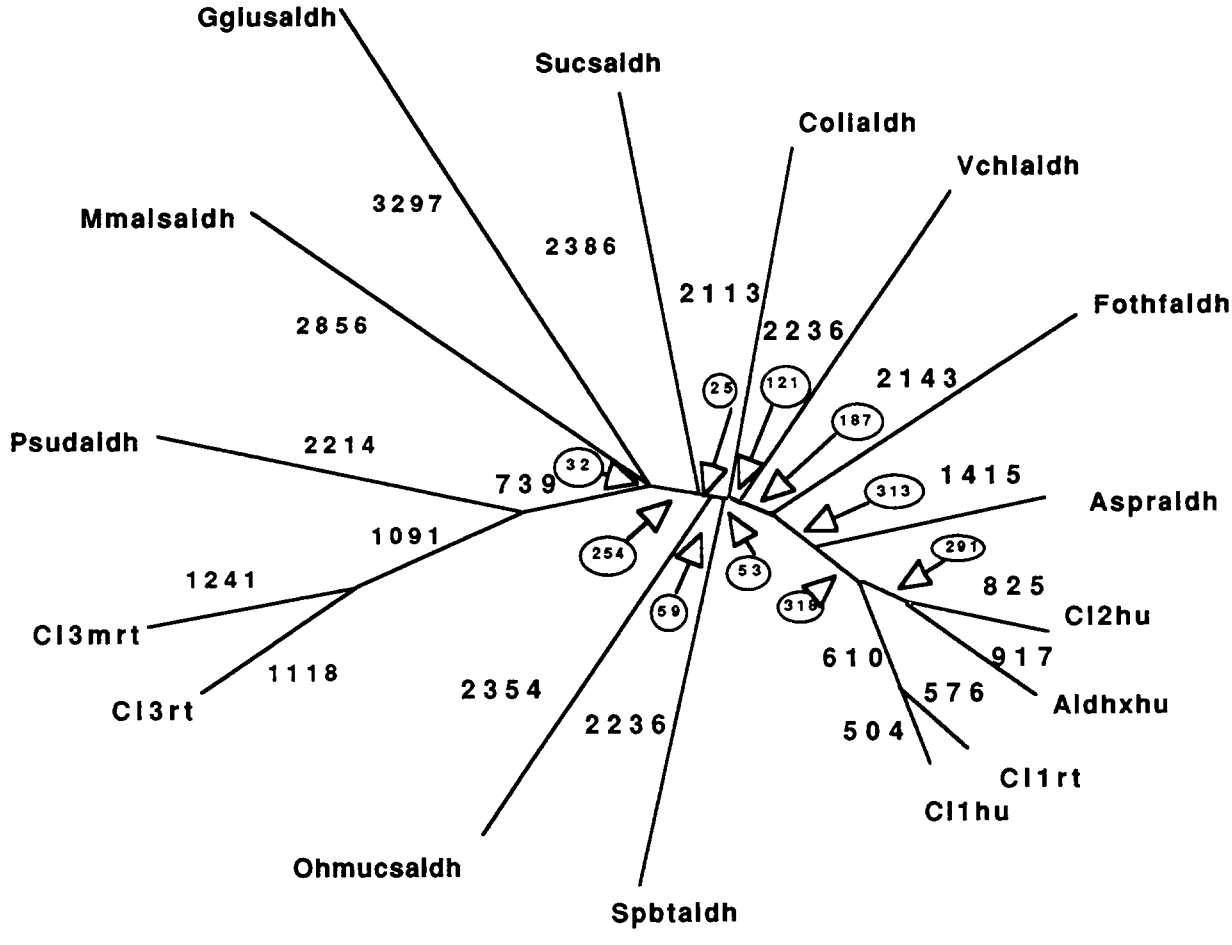

Fig. 4. Unrooted evolutionary tree derived from the sequences as aligned in Figure 1. The tree was constructed using the FITCH program in the PHYLIP suite of programs (Fitch & Margoliash, 1967; Felsenstein, 1990) and is drawn to the scale of the distances given to the side of each branch. Sequence abbreviations are as given in the legend to Figure 1.

the tendency to maintain this relationship in all but the two class 3 AlDHs (both positions uncharged), *Pseudomonas* and methylmalonyl semialdehyde dehydrogenase (each with lone arginine residues), and hydroxymuconic semialdehyde dehydrogenase (with double basic residues).

Clearly, our understanding of AlDH structure–function relationships would be advanced by the availability of a tertiary structural model that is determined experimentally. Several groups are now working separately on crystals of the class 1, 2, and 3 enzymes, yet despite some progress (Rose et al., 1991; Hurley & Weiner, 1992) none have advanced to the stage of having suitable heavy-atom derivatives. Thus, for now, the above observations gleaned from sequence alignments must guide our efforts at site-directed mutagenesis, which has most recently been aimed at development of better heavy-atom derivatives.

Methods

The AlDH sequences aligned were from human liver cytosol (class 1) (Hempel et al., 1984), phenobarbital induction of rat liver (Dunn et al., 1989), human liver mitochondria (class 2) (Hempel et al., 1985), "AlDHx" from human genomic library screening (Hsu & Chang, 1991), *Aspergillus* (Pickett et al., 1987), spinach (a betaine aldehyde dehydrogenase) (Weretilnyk & Hanson, 1990), TCDD induction of rat liver (class 3) (Hempel et al., 1989) and *Pseudomonas* (Kok et al., 1989), succinyl semialdehyde dehydrogenase from *E. coli* (Niegmann et al., 1992), methyl malonic semialdehyde dehydrogenase from rat liver (Kedishvilli et al., 1992), membrane-bound AlDH from rat liver microsomes (another class 3 type) (Miyauchi et al., 1991), and soluble AlDHs from *Vibrio cholerae* (Parsot & Mekalanos, 1991) and *E. coli* (Heim & Strehler, 1991). A sequence reported from yeast (Saigal et al., 1991) has been omitted because certain portions of that sequence have been questioned by these authors.

Substrate-specific AlDHs include γ-glutamyl semialdehyde dehydrogenase from yeast (Krzywicki & Brandriss, 1984), hydroxymuconic semialdehyde dehydrogenase from *Pseudomonas* (Norlund & Shingler, 1990), and the C-terminal half of formyl tetrahydrofolate dehydroge-

nase, a chimeric protein (Cook et al., 1991). The latter represents the first example of a chimeric enzyme composed of an aldehyde dehydrogenase domain joined to other functional domains in a single polypeptide.

The sequences were obtained from the NBRF data base (George et al., 1986) or from the original sources. An initial multiple alignment was made using the CGC Pileup program (Genetics Computer Group, 1991), and refinements, primarily at the N- and C-terminal ends, were introduced manually using the GCG Lineup program. The alignment display (Fig. 1) was generated using the MALIGNED editor and MALFORMED formatter (Clark, 1992). Secondary structural assessments were made on the basis of Chou and Fasman (1978) numerical potentials as refined by Ralph et al. (1987). In order to avoid overrepresenting the contributions of the closely similar class 1, 2, and 3 sequence pairs, the rat class 1, human AlDHx, and rat microsomal sequences were omitted from the alignment used for calculation of average Chou–Fasman values. These values and standard deviations were compiled using four-residue windows for helical and turn predictions, and three-residue windows for β-strand predictions. The averages are centered on the second residue of the window for both helices and strands, and on the first residue for reverse turns; values flanking gaps of individual sequences were averaged across the gaps. Graphical compilations of the data were made using CricketGraph on a Macintosh computer.

Acknowledgments

Funding from the National Institute on Alcohol Abuse and Alcoholism (AA06985, to J.H. and R.L.) and the NIH Division of Research Resources (cooperative agreement RR06009, to the Pittsburgh Supercomputing Center) is acknowledged.

References

Abola, E., Bernstein, F.C., Bryant, S.H., Koetzel, T.F., & Weng, J. (1987). Protein Data Bank. In *Crystallographic Databases—Information Content, Software Systems, Scientific Applications* (Allen, F.H., Bergerhoff, G., & Sievers, R., Eds.), pp. 107–132. Data Communications of the International Union of Crystallography, Bonn, Germany.

Abriola, D.P., Fields, R., Stein, S., MacKerell, A.D., & Pietruszko, R. (1987). Active site of human liver aldehyde dehydrogenase. *Biochemistry 26*, 5679–5684.

Argos, P. (1989). Predictions of protein structure from gene and amino acid sequences. In *Protein Structure: A Practical Approach* (Creighton, T.E., Ed.), pp. 169–190. IRL Press, Oxford, UK.

Bernstein, F.C., Koetzel, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Borrás, T., Persson, B., & Jörnvall, H. (1989). Eye lens ζ-crystallin: Relationships to the family of "long-chain" alcohol/polyol dehydrogenases. *Biochemistry 28*, 6133–6139.

Brändén, C. & Tooze, J. (1991). *Introduction to Protein Structure*. Garland, New York.

Chou, P.Y. & Fasman, G.D. (1978). Empirical predictions of protein conformations. *Annu. Rev. Biochem. 47*, 251–276.

Clark, S.P. (1992). Maligned: A multiple sequence editor. *CABIOS 8*, 535–538.

Cook, R.J., Lloyd, R.S., & Wagner, C. (1991). Isolation and characterization of cDNA clones for rat liver 10-formyl- tetrahydrofolate dehydrogenase. *J. Biol. Chem. 266*, 4965–4973.

Cooper, D.L. & Baptist, E.W. (1991). Degenerate oligonucleotide cloning of the BCP54/AlDH3 cDNA. *PCR Methods Applic. 1*, 57–262.

Crabb, D.W., Edenberg, H.J., Bosron, W.F., & Li, T.-K. (1989). Genotypes of aldehyde dehydrogenase deficiency and alcohol sensitivity. The inactive $ALDH_{22}$ allele is dominant. *J. Clin. Invest. 83*, 314–316.

Creighton, T.E. (1993). *Proteins: Structures and Molecular Properties*. W.H. Freeman, New York.

Dayhoff, M.O., Barker, W.C., & Hunt, L.T. (1983). Establishing homologies in protein sequences. *Methods Enzymol. 91*, 524–545.

Doolittle, R.F. (1989). Redundancies in protein sequences. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), p. 599–623. Plenum, New York.

Dunn, T.J., Koleske, A.J., Lindahl, R., & Pitot, H.C. (1989). Phenobarbital-inducible aldehyde dehydrogenase in the rat. *J. Biol. Chem. 264*, 13057–13065.

Eklund, H., Nordström, B., Zeppezauer, E., Söderlund, G., Ohlsson, I., Boiwe, T., Söderberg, B.O., Tapia, O., Brändén, C.I., & Åkesson, Å. (1976). Three-dimensional structure of horse liver alcohol dehydrogenase at 2.4 Ångstroms resolution. *J. Mol. Biol. 102*, 27–59.

Felsenstein, J. (1990). *PHYLIP Manual*, Version 3.3. University Herbarium, University of California, Berkeley.

Fetrow, J. & Rose, G.D. (1990). Loops in globular proteins. In *Protein Folding* (Gierasch, L.M. & King, J., Eds.), pp. 18–28. American Association for the Advancement of Science, Washington, D.C.

Fitch, W.M. & Margoliash, E. (1967). Construction of phylogenetic trees. *Science 155*, 279–284.

Genetics Computer Group. (1991). *Program Manual for the GCG Package*, Version 7, April 1991. Genetics Computer Group, Madison, Wisconsin.

George, D.G., Barker, W.C., & Hunt, L.T. (1986). The Protein Identification Resource. *Nucleic Acids Res. 14*, 11–15.

Heim, R. & Strehler, E.E. (1991). Cloning an *Escherichia coli* gene encoding a protein remarkably similar to mammalian aldehyde dehydrogenases. *Gene 99*, 15–23.

Hempel, J., Harper, K., & Lindahl, R. (1989). Inducible class 3 aldehyde dehydrogenase from rat hepatocellular carcinoma and 2,3,7,8-tetrachlorodibenzo-*p*-dioxin-treated liver: Distant relationship to the class 1 and 2 enzymes from mammalian liver cytosol/mitochondria. *Biochemistry 28*, 1160–1167.

Hempel, J., Kaiser, R., & Jörnvall, H. (1985). Mitochondrial aldehyde dehydrogenase from human liver: Primary structure, differences in relation to the cytosolic enzyme and functional correlations. *Eur. J. Biochem. 153*, 13–28.

Hempel, J., Nicholas, H., & Jörnvall, H. (1991). Thiol proteases and aldehyde dehydrogenases: Evolution from a common thiolesterase precursor? *Proteins Struct. Funct. Genet. 11*, 176–183.

Hempel, J., Pietruszko, R., Fietzek, P., & Jörnvall, H. (1982). Identification of a segment containing a reactive cysteine residue in human liver aldehyde dehydrogenase. *Biochemistry 21*, 6834–6838.

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, L.C., Bendel, R.E., & Yoshida, A. (1988). Genomic structure of the human mitochondrial aldehyde dehydrogenase gene. *Genomics 2*, 57–65.

Hsu, L.C. & Chang, W.-C. (1991). Cloning and characterization of a new functional human aldehyde dehydrogenase gene. *J. Biol. Chem. 266*, 12257–12265.

Hsu, L.C., Chang, W.-C., & Yoshida, A. (1989). Genomic structure of the human cytosolic aldehyde dehydrogenase gene. *Genomics 5*, 857–865.

Hurley, T.D. & Weiner, H. (1992). Crystallization and preliminary X-ray analysis of bovine liver mitochondrial AlDH. *J. Mol. Biol. 227*, 1255–1257.

Johansson, J., von Bahr-Lindström, H., Jeck, R., Woenckhaus, C., & Jörnvall, H. (1988). Mitochondrial aldehyde dehydrogenase from horse liver: Correlations of the same species variants for both the cytosolic and the mitochondrial forms of an enzyme. *Eur. J. Biochem. 172*, 527–533.

Jones, D.T., Taylor, W.R., & Thornton, J.M. (1992). The rapid gener-

ation of mutation data matrices from protein sequences. *CABIOS 8*, 275–282.

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

Kedishvilli, N.Y., Popov, K.M., Rougraff, P.M., Zhao, Y., Crabb, D.W., & Harris, R.A. (1992). CoA-dependent methylmalonate semialdehyde dehydrogenase, a unique member of the aldehyde dehydrogenase superfamily. *J. Biol. Chem. 267*, 19724–19729.

Kok, M., Oldenhuis, R., vander Linden, M.P.G., Meulenberg, C.H.C., Kingma, J., & Witholt, B. (1989). The *Pseudomonas oleovirans alkBAC* operon encodes two structurally related rubredoxins and an aldehyde dehydrogenase. *J. Biol. Chem. 264*, 5442–5451.

Krzywicki, K.A. & Brandriss, M.C. (1984). Primary structure of the nuclear PUT2 gene involved in the mitochondrial pathway for proline utilization in *Saccharomyces cerevisiae*. *Mol. Cell. Biol. 4*, 2837–2842.

Kurys, G., Shah, P., Reed, D., Ambroziak, W., & Pietruszko, R. (1992). Human aldehyde dehydrogenase third isoenzyme: cDNA cloning and deduced amino acid sequence. *FASEB J. 6*, A77 [Abstr.].

Lesk, A.M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: The structures and evolutionary dynamics of the globins. *J. Mol. Biol. 136*, 225–270.

Lindahl, R. & Hempel, J. (1991). Aldehyde dehydrogenases: What can be learned from a baker's dozen sequences? *Adv. Exp. Med. Biol. 284*, 1–8.

Loomes, K. & Jörnvall, H. (1991). Structural organization of aldehyde dehydrogenases probed by limited proteolysis. *Biochemistry 30*, 8865–8870.

McCaffery, P., Tempst, P., Lara, G., & Drager, U.C. (1991). Aldehyde dehydrogenase is a positional marker in the retina. *Development 112*, 693–702.

Miyauchi, K., Masaki, R., Taketani, S., Yamamoto, A., Akayama, M., & Tashiro, Y. (1991). Molecular cloning, sequencing, and expression of cDNA for rat liver microsomal aldehyde dehydrogenase. *J. Biol. Chem. 266*, 19536–19542.

Neurath, H. (1943). The role of glycine in protein structure. *J. Am. Chem. Soc. 65*, 2039–2040.

Niegmann, E., Schulz, A., & Bartsch, K. (1992). Sequence of *E.coli* succinic semialdehyde dehydrogenase. Submission to SwissProt database (accession Gabd_Ecoli).

Norlund, I. & Shingler, V. (1990). Nucleotide sequences of the metacleavage pathway enzymes: 2-Hydroxymuconic semialdehyde dehydrogenase and 2-hydroxymuconic semialdehyde hydrolase from *Pseudomonas* CF600. *Biochim. Biophys. Acta 1049*, 227–230.

Parsot, C. & Mekalanos, J.J. (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.

Pereira, F., Rosenmann, E., Nylen, E., Kaufman, M., Pinsky, L., & Wrogemann, K. (1991). The 56 kDa androgen binding protein is an aldehyde dehydrogenase. *Biochem. Biophys. Res. Commun. 175*, 831–838.

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

Pickett, M., Gwynne, D.I., Buxton, F.P., Elliott, R., Davies, R.W., Lockington, R.A., Scazzocchio, C., & Sealy-Lewis, H.M. (1987). Cloning and characterization of the *ald A* gene of *Aspergillus nidulans*. *Gene 51*, 217–226.

Pietruszko, R., Blatter, E., Abriola, D.P., & Prestwich, G. (1991). Localization of cysteine 302 at the active site of aldehyde dehydrogenase. *Adv. Exp. Med. Biol. 284*, 19–30.

Ralph, W.R., Webster, T., & Smith, T.F. (1987). A modified Chou and Fasman protein structure algorithm. *Comput. Applic. Biosci. 3*, 211–216.

Richardson, J.S. & Richardson, D.C. (1987). Some design principles: Betabellin. In *Protein Engineering* (Oxender, D.L. & Fox, C.F., Eds.), pp. 149–163. Liss, New York.

Richardson, J.S. & Richardson, D.C. (1989). Principles and patterns of protein conformation. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), pp. 1–98. Plenum, New York.

Rondeau, J.-M., Tête-Favier, F., Podjarny, A., Reymann, J.-M., Barth, P., Biellmann, J.-F., & Moras, D. (1992). Novel NADPH-binding domain revealed by the crystal structure of aldose reductase. *Nature 355*, 469–472.

Rose, J., Hempel, J., Kuo, I., Lindahl, R., & Wang, B.-C. (1991). Preliminary crystallographic analysis of class 3 rat liver aldehyde dehydrogenase. *Proteins Struct. Funct. Genet. 8*, 305–308.

Rossmann, M.G., Moras, D., & Olsen, K.W. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature 250*, 194–199.

Saigal, D., Cunningham, S.J., Farrés, J., & Weiner, H. (1991). Molecular cloning of the mitochondrial aldehyde dehydrogenase gene of *Saccharomyces cerevisiae* by genetic complementation. *J. Bacteriol. 173*, 3199–3208.

Takahashi, K., Weiner, H., & Filmer, D.L. (1981). Effects of pH on horse liver aldehyde dehydrogenase: Alterations in metal ion activation, number of functioning active sites, and hydrolysis of the acyl intermediate. *Biochemistry 21*, 6225–6230.

von Bahr-Lindström, H., Hempel, J., & Jörnvall, H. (1984). The cytoplasmic isoenzyme of horse liver aldehyde dehydrogenase. *Eur. J. Biochem. 141*, 37–42.

von Bahr-Lindström, H., Jeck, R., Woenckhaus, C., Sohn, S., Hempel, J., & Jörnvall, H. (1985). Characterization of the coenzyme-binding site of liver aldehyde dehydrogenase: Differential reactivity of coenzyme analogs. *Biochemistry 24*, 5847–5851.

Weretilnyk, E.A. & Hanson, A.D. (1990). Molecular cloning of a plant betaine-aldehyde dehydrogenase, an enzyme implicated in adaptation to salinity and drought. *Proc. Natl. Acad. Sci. USA 87*, 2745–2749.

Wierenga, R.K. & Hol, W.G.J. (1983). Predicted nucleotide-binding properties of p21 and its cancer-associated variant. *Nature 302*, 842–844.