FOR THE RECORD

Ethanol utilization regulatory protein: Profile alignments give no evidence of origin through aldehyde and alcohol dehydrogenase gene fusion

HUGH B. NICHOLAS, JR.,¹ BENGT PERSSON,² HANS JÖRNVALL,² AND JOHN HEMPEL³

¹ Pittsburgh Supercomputing Center, Pittsburgh, Pennsylvania 15213

² Karolinska Institute, Department of Medical Biochemistry and Biophysics, S-171 77 Stockholm, Sweden

³ University of Pittsburgh Biotechnology Center, Department of Molecular Genetics & Biochemistry, Pittsburgh, Pennsylvania 15219

(RECEIVED August 9, 1995; ACCEPTED September 29, 1995)

Abstract: The suggestion that the ethanol regulatory protein from *Aspergillus* has its evolutionary origin in a gene fusion between aldehyde and alcohol dehydrogenase genes (Hawkins AR, Lamb HK, Radford A, Moore JD, 1994, *Gene* 146:145–158) has been tested by profile analysis with aldehyde and alcohol dehydrogenase family profiles. We show that the degree and kind of similarity observed between these profiles and the ethanol regulatory protein sequence is that expected from random sequences of the same composition. This level of similarity fails to support the suggested gene fusion.

Keywords: alcohol dehydrogenase; aldehyde dehydrogenase; ethanol utilization regulatory protein; profile analysis; sequence conservation; sequence similarity

Present-day proteins are often derived from domains and ancestral building units of smaller size, reflecting primitive proteins (Doolittle, 1979). They, in turn, may have originated from different families, hence giving multiple superfamily assignments to the present-day forms. Multifunctional proteins in particular, such as regulatory proteins that bind two or more different molecules, are thought to generally arise by the fusion of genes originally encoding separate functions.

Because of distant relationships, great care is necessary in evaluation of sequence similarities (Doolittle, 1979). Apparent sequence similarity may be insignificant for a number of reasons (Henikoff, 1991; Altschul et al., 1994). Evidence supporting the possibility that two proteins (or the genes encoding them) are homologous frequently consists of a catalogue of similarities between the proteins based on their amino acid sequence,

catalytic activity, substrate, cofactor or effector binding, or other biochemical property. Levels of sequence similarity that are marginal (in the “Twilight Zone” [Doolittle, 1986]) can be enhanced by showing that the observed similar or identical elements are essential for biochemical function or for structural integrity of the protein. We have used this approach to find evidence of relationships between aldehyde dehydrogenase (ALDH) and thiol proteases (Hempel et al., 1990), and between alcohol dehydrogenase (ADH) and a variety of other enzymes (Persson et al., 1993, 1994; Jörnvall et al., 1995).

The ethanol metabolism-regulating protein from *Aspergillus* (AlcR) was recently hypothesized to have resulted from fusion of the ALDH and ADH genes (Hawkins et al., 1994). In the alignment of these two enzymes with the regulating protein, 111 matching residues were found among 750 aligned positions, corresponding to ~15% identity. This level of similarity seems high enough to prompt scrutiny for additional supporting features. We have, therefore, examined the similarity between AlcR and the ALDH or ADH families, but find insufficient evidence to support the relationships previously claimed.

Results and discussion: We first used profiles (Gribskov et al., 1990) derived from multiple sequence alignments of ALDH (Hempel et al., 1993) and ADH (Persson et al., 1994) to investigate the sequence similarity between these two enzymes and AlcR. Profiles provide a sensitive test of the similarity between a sequence family and a proposed homologue by recording the pattern of amino acids that have been conserved during evolution of the family. Thus, the proposed homologue is compared to this pattern of conservation rather than to a single idiosyncratic sequence from the family, and conserved residues are given added importance in determining similarity.

The best-scoring alignments of our profiles with the AlcR protein (included in the Electronic Appendix) are incompatible with the alignment suggested previously (Hawkins et al., 1994), which

Reprint requests to: Hugh B. Nicholas Jr., Pittsburgh Supercomputing Center, 4400 Fifth Avenue, Pittsburgh, Pennsylvania 15213; e-mail: nicholas@psc.edu.

aligned ALDH with the N-terminal part, and ADH with the C-terminal part of AlcR. In contrast, our best similarity score now suggests, if anything, the opposite concatenation – with the ADH profile possibly aligned with the N-terminal end of AlcR and the ALDH profile with the C-terminal part (Fig. 1). However, this optimal case only indicated weak similarity by aligning relatively short parts of each profile with the AlcR protein (Fig. 1). Aligning our profiles with 999 randomized AlcR sequences allowed comparison of all 1,000 (native plus randomized) highest scores. The tallied scores yield a significance probability (Table 1) that shows the observed similarity is too weak to support the hypothesis of a gene fusion.

In order to insure that we did not inadvertently miss a significant alignment by selecting inappropriate conditions, we tried several alignment strategies. First, we computed a traditional Smith and Waterman (1981) local alignment originally proposed for use with profiles. For this alignment, we adjusted the New-gap and Gap penalties until the gaps were similar in size and number to the gaps in the previously reported alignments. Second, with the same gap settings, we used a feature available in our profile alignment program (Ropelewski AJ, Nicholas HB Jr, and Gribskov M, in prep.) to force the entire profile to align with the AlcR sequence (Sellers, 1979). Third, we used NewGap and Gap penalties large enough to prevent any gaps from being introduced into the alignment. This yields a maximal seg-

Table 1. Alignment of ALDH and ADH profiles with the AlcR protein^a

Profile	Penalty for		Alignment type	Best score vs. AlcR	Significance probability
	Gap initiation	Gap elongation			
ALDH	-10.0	-1.0	Local	185.15	0.210
ALDH	-10.0	-1.0	Global	168.22	0.453
ALDH	-75.0	-30.0	Local	180.30	0.601
ADH	-10.0	-1.0	Local	103.89	0.811
ADH	-10.0	-1.0	Global	98.14	0.468
ADH	-75.0	-30.0	Local	99.09	0.499

^a Significance probability is the fraction of 999 randomized AlcR sequences plus the native sequence scoring as high as the native AlcR sequence. Actual randomized sequences are available by e-mail to H.B.N.

ment pair alignment similar to alignments produced by the BLAST program (Altschul et al., 1990). None of these three different alignment strategies yielded a similarity great enough to support the hypothesis that the AlcR protein is homologous to either ALDH or ADH (Table 1).

Fig. 1. Diagram of the optimal alignment of the *Aspergillus* AlcR protein sequence with the profiles from ADH and ALDH families, contrasted to the alignment of the *Aspergillus* AlcR protein sequence with the *Aspergillus* ADH and ALDH sequences. The solid black bar in the center of the diagram represents the *Aspergillus* AlcR protein sequence. Above the solid black bar are two bars representing the ADH and ALDH profiles as aligned to the *Aspergillus* AlcR protein sequence (this study). Below the solid black bar are two bars representing the *Aspergillus* ADH and ALDH protein sequences as previously aligned to the *Aspergillus* AlcR protein sequence (Hawkins et al., 1994). Note that the striped bar is associated with the ADH family, whether a sequence or a profile is represented and the gray bar is associated with the ALDH family in the same way. Enhanced emphasis on conserved residues obtained through the use of profiles reverses the order in which ADH and ALDH align with AlcR. Numbers along the protein sequence schematics refer to actual position numbers. Numbers along the profile schematics refer to positions within the multiple sequence alignment of the protein family and begin at the first position of the multiple sequence alignment used in the profile. These multiple sequence alignments are in the Electronic Appendix and the positions used in the profiles are marked.

Inspection of the previous alignment (Hawkins et al., 1994) for matches with the strictly conserved residues in the separate ADH and ALDH family sequence alignments (Electronic Appendix) reveals that only 5 of 33 residues (15%) are conserved. This is the same as the overall level of identity suggested by the pairwise alignment. A broader view of the important elements in the two protein families was obtained from the profiles, assigning weight to each position in proportion to the observed extent of conservation. The ADH and ALDH profiles were found to yield better indices of similarity when aligned in opposite order (Fig. 1) relative to the previous alignment (Hawkins et al., 1994). Further, as the results of comparisons with randomized AlcR sequences show (Table 1), the similarity between the AlcR sequence and the profiles is typical of that expected between the profiles and a random sequence of amino acids with the same amino acid composition as AlcR.

Table 2 shows that the patterns of conserved residues in the previous alignment more closely resemble the average amino acid compositions of the aligned sequence pairs and are distinctly dissimilar from the patterns of consensus residues in the two families (Electronic Appendix). In particular, in both the ADH and ALDH families, there is a strong preference for conservation of Gly, reflecting the overriding evolutionary constraint of maintaining folding patterns in the tertiary structure (Jörnvall et al., 1987; Hempel et al., 1993). Because these insights have emerged from alignments including the most distantly related members of the respective families, with some pairs clearly in the "Twilight Zone" of about 15% positional identity, the same trends would be expected to apply to the proposed fusion product

(Hawkins et al., 1994) with similar levels of identity. The previous alignment, rather than maintaining many of the completely conserved amino acids of the ALDH and ADH families, instead shows a pattern of conservation mimicking the amino acid composition of the proteins.

We compared the AlcR protein with all three sections of the PIR protein database using the Smith-Waterman algorithm as a final check against the possibility that the AlcR protein showed noteworthy similarity to any ADH or ALDH protein. A few sequences, notably zinc-binding proteins, scored between 70 and 86, which is compatible with the observation by Hawkins et al. (1994) that the AlcR protein contains a possible zinc-binding site (Kulmberg et al., 1991). The bulk of the database scores are below 60. The highest ADH score was 41, with 4,253 other sequences scoring as high or higher. The highest ALDH score was 46, with 1,120 other sequences scoring as high or higher.

Thus, the use of objective evaluations of the previously suggested sequence similarity between the ADH and ALDH sequences and the AlcR sequence confirms that the level of similarity observed is too low to support the fusion hypothesis. From all the facts, we fail to find any support for an ancestral alcohol and aldehyde dehydrogenase gene fusion to explain the origin of the alcohol regulatory protein.

Although the relationship reported by Hawkins et al. (1994) has its origin in a CLUSTAL V analysis, that analysis also contains the unverified assumption that the three sequences analyzed are homologous. An analysis such as the one presented here is a necessary adjunct to verify the assumption of homologous sequences. This illustrates the danger of being led by an intuitively

Table 2. Residue conservation patterns

Amino acid residue	Aldehyde dehydrogenase			Alcohol dehydrogenase		
	AlcR/ALDH Hawkins alignment	Expected ^a from total composition	Conserved in ALDH family	AlcR/ADH Hawkins alignment	Expected ^a from total composition	Conserved in ADH family
A	10	6	—	8	6	—
C	—	1	1	1	1	1
D	6	3	—	5	4	—
E	4	3	1	2	3	1
F	2	2	1	—	1	—
G	6	3	11	2	5	7
H	—	1	—	—	2	1
I	3	3	—	1	3	—
K	—	2	1	2	3	—
L	2	4	—	10	6	—
M	—	1	—	—	1	—
N	—	3	2	—	—	—
P	2	3	3	4	3	—
Q	2	2	—	3	3	—
R	4	3	1	2	3	—
S	6	4	1	4	4	—
T	5	4	1	3	2	—
V	1	3	—	8	5	—
W	—	1	—	—	1	—
Y	—	1	—	3	2	—
Sum	53	53	23	58	58	10

^a Expected values are taken from the combined *Aspergillus* AlcR and ALDH or ADH sequences, and are the products of the mole percent of each residue normalized to total 53 or 58, and rounded to the nearest integer.

plausible relationship without ever subjecting the alignment to objective scrutiny. Testing any scientific hypothesis requires some control for false positive results. The randomization test we performed with the ADH and ALDH profiles is a rigorous statistical test of the gene fusion hypothesis because the hypothesis was derived from the organization of the genes on the chromosome and not from a database search. Database test results also provide an objective basis for assessing the similarity observed between putatively homologous sequences.

Methods: For the ALDH family, we used a 16-sequence alignment (Hempel et al., 1993). We trimmed the ends of the alignment to focus on the core present in all ALDH family members. Our core begins with the partially conserved tryptophan at index position 119 of the alignment, continuing through index 563, with one minor change. We removed the two gaps following the conserved serine at index 557 in the yeast gamma-glutamyl semi-aldehyde dehydrogenase to produce the sequence SGTGKAG and then, at the Asp in this segment, introduced a single residue gap into all the other sequences in the alignment. This change enhances the degree of conservation in this region, while reducing the number of insertion or deletion events postulated by the alignment.

For the ADH family, we used an alignment of 46 characterized medium-chain alcohol dehydrogenases based on an earlier alignment (Jörnvall et al., 1987). We included the amino acids from one position before the conserved Cys at position 46 through one residue past the partially conserved Gly at position 320. Both profiles were constructed using the PAM 250 similarity scores, a set of similarity scores appropriate for the degree of conservation exhibited by the alignment of Hawkins et al. The ADH and the ALDH multiple sequence alignments are included in the Electronic Appendix.

Randomized AlcR sequences were generated by shuffling the amino acids in the sequence so that each amino acid was moved a minimum of three times and an average of six times from its position in the previously generated random sequence.

Supplementary material on Electronic Appendix: The Electronic Appendix contains three files for each of the two protein families, ALDH and ADH, used as data in this study (six files total). The first file for each protein family is the multiple sequence alignment in machine-readable ASCII text. The second file is a postscript file with the conserved regions in each alignment highlighted. The third file for each protein family is the align-

ment of the profile with ethanol metabolism-regulating protein sequence. A seventh file describes the first six files in more detail.

Acknowledgments: We thank Sidney Morris for his keen observations. Support from the NIH Division of Research Resources (grant RR06009 to the Pittsburgh Supercomputing Center), AA06985 and the Swedish Medical Research Council is gratefully acknowledged.

References

- Altschul SF, Boguski MS, Gish W, Wootton JC. 1994. Issues in searching molecular sequence databases. *Nature Genetics* 6:119–129.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410.
- Doolittle RF. 1979. Protein evolution. In: Neurath H, Hill RH, eds. *The proteins, 3rd ed, vol 4*. New York: Academic Press. pp 1–118.
- Doolittle RF. 1986. *Of URFs and ORFs: A primer on how to analyze derived amino acid sequences*. Mill Valley, California: University Science Books.
- Gribskov M, Luthy R, Eisenberg D. 1990. Profile analysis. *Methods Enzymol* 183:146–162.
- Hawkins AR, Lamb HK, Radford A, Moore JD. 1994. Evolution of transcription-regulating proteins by enzyme recruitment: Molecular models for nitrogen metabolite repression and ethanol utilization in eukaryotes. *Gene* 146:145–158.
- Hempel J, Nicholas H, Jörnvall H. 1990. Aldehyde dehydrogenases and thiol proteases: Evolution from a common thioesterase ancestor? *Proteins Struct Funct Genet* 11:176–183.
- Hempel J, Nicholas H, Lindahl R. 1993. Aldehyde dehydrogenases: Widespread structural and functional diversity within a shared framework. *Protein Sci* 2:1890–1900.
- Henikoff S. 1991. Playing with blocks: Some pitfalls of forcing multiple alignments. *New Biologist* 3:1148–1154.
- Higgins DG, Bleasby AJ, Fuchs R. 1992. CLUSTAL V: Improved software for multiple sequence alignment. *CABIOS* 8:189–191.
- Jörnvall H, Persson B, Jeffery J. 1987. Characteristics of alcohol/polyol dehydrogenases. *Eur J Biochem* 167:195–201.
- Jörnvall H, Persson B, Krook M, Atrain S, González-Duarte R, Jeffery J, Ghosh D. 1995. Short-chain dehydrogenases/reductases (SDR). *Biochemistry* 34:6003–6013.
- Kulmberg P, Prange T, Mathieu M, Sequeval D, Scazzocchio C, Felenbok B. 1991. Correct intron splicing generates a new type of putative zinc binding domain in a transcriptional activator of *Aspergillus*. *FEBS Lett* 280:11–16.
- Persson B, Hallborn J, Walfridsson M, Hahn-Hägerdal B, Keränen S, Penttilä M, Jörnvall H. 1993. Dual relationships of xylitol and alcohol dehydrogenases in families of two protein types. *FEBS Lett* 324:9–14.
- Persson B, Zigler JS Jr, Jörnvall H. 1994. A super-family of medium-chain dehydrogenases/reductases (MDR). Sub-lines including zeta-crystallin, alcohol and polyol dehydrogenases, quinone oxidoreductases, enoyl reductases, VAT-1 and other proteins. *Eur J Biochem* 226:15–22.
- Sellers PH. 1979. Pattern recognition in genetic sequences. *Proc Natl Acad Sci USA* 76:3041.
- Smith TF, Waterman MS. 1981. Identification of common molecular sub-sequences. *J Mol Biol* 147:195–197.