Intron-dependent evolution of the nucleotide-binding domains within alcohol dehydrogenase and related enzymes

Gregg Duester[§], Hans Jörnvall⁺ and G.Wesley Hatfield

Department of Microbiology and Molecular Genetics, University of California, Irvine, CA 92717, USA and ⁺Department of Chemistry I, Karolinska Institutet, S-104 01-Stockholm, Sweden

Received 11 December 1985; Revised and Accepted 3 February 1986

ABSTRACT

It has been suggested that the intron/exon structure of a gene corresponds to its evolutionary history. Accordingly, early in evolution DNA segments encoding short functional polypeptides may have been rearranged (exon-shuffling) to create full-length genes and RNA splicing may have been developed to remove intervening sequences (introns) in order to preserve translational reading frames. A conflicting viewpoint would be that introns were randomly inserted into previously uninterrupted genes after their initial evolutionary development. If so, the sites of introns would be unlikely to consistently reflect the domain structure of the protein. To address this question, the intron/exon structure of the gene encoding human alcohol dehydrogenase (ADH) was determined and compared to the gene structures for other ADHs and related proteins, all of which possess nucleotide-binding domains. Our results indicate that the introns in the nucleotide-binding domains of all the genes examined do indeed fall at positions which separate the short functional polypeptides (i.e. β strands) which are believed to comprise this domain. We argue that our data is most easily explained by the hypothesis that introns were present in an ancestral nucleotide-binding domain which was later rearranged by exon-shuffling to form the various dehydrogenases and kinases which utilize such a domain.

INTRODUCTION

Since the discovery that most genes in higher eukaryotes are broken up into short coding segments (exons) interrupted by noncoding segments (introns), the evolutionary origin of introns has remained elusive. Since introns in some eukaryotic genes have been observed to fall between functional protein domains it has been suggested that introns participated in protein evolution by facilitating the recombination of exons, i.e. exon-shuffling (1). This hypothesis was extended to suggest that introns were present in the genes of primitive organisms and were the result of the initial evolutionary development of full-length genes from small exonic units encoding functional polypeptides (2,3). According to this view bacteria and lower eukaryotes lost most of their introns later in evolution. Another view suggests that introns are the result of insertions of transposable elements into previously uninterrupted genes (4,5). Both viewpoints have recently been reviewed (6). Since the two viewpoints are not mutually exclusive it is possible that

some introns observed in a given gene were present in its primordial ancestor and that other introns were inserted later in evolution.

To address this issue we have examined the intron/exon organization of human liver alcohol dehydrogenase (ADH) to ascertain if the positions of any introns correlate with the junctions of functional domains known for the horse enzyme (7). Previous analyses of the maize ADH gene indicated a correlation (8,9). In addition, since ADH is conserved among animals, plants, yeast, and bacteria (9-20), a comparison of the intron/exon structures in various ADH genes may shed light on the evolutionary origin of introns. Our DNA sequence analysis of the human β ADH gene revealed the presence of eight introns (21), three of which separate short functional regions within the nucleotide-binding domain. A comparison of the sites of the human ADH introns with those of other ADH genes (9,22,23) has revealed interesting correlations. Additionally, a comparison of the intron/exon structures of ADH genes with those of other genes which possess a nucleotide-binding domain (24-27) has revealed a very clear relationship between the locations of certain introns and the secondary structure of nucleotide-binding domains. Collectively, the results support the view that some introns in ADH genes and related genes were incorporated when the genes were originally constructed.

MATERIALS AND METHODS

Amino Acid Sequence Alignments.

The amino acid sequence for human β ADH predicted from the nucleotide sequence was aligned with the amino acid sequences for horse liver ADH, yeast ADH, and sheep liver sorbitol dehydrogenase (SDH) using the conformation-related alignment of Eklund et al. (12). The human β ADH amino acid sequence was also aligned with the sequence of maize ADH1 (9) and the amino terminal sequence of *Bacillus stearothermophilus* ADH (10, 13).

Comparison of the Intron/Exon Structures of Nucleotide-binding Domains.

The intron/exon structure of the nucleotide-binding domain within human β ADH (21) was compared to the intron/exon structures of the corresponding domains in maize ADH (9, 22), *Drosophila* ADH (23, 28), chicken glyceraldehyde phosphate dehydrogenase (GAPDH) (24), mouse lactate dehydrogenase (LDH-A) (26), and chicken pyruvate kinase (PK) (27). The amino acid residues comprising the conserved β strands which are present in the nucleotide-binding domains of the enzymes described above were deduced by alignment of their amino acid sequences with those of homologous enzymes which have been subjected to x-ray crystallographic analysis and for which the exact locations of the β strands within the nucleotide-binding domains are known; i.e. horse liver ADH (7), lobster GAPDH (29), dogfish LDH (29), and cat muscle PK (30).

RESULTS

Intron/Exon Structure of Human ADH.

To determine the intron/exon structure for human ADH we isolated a cDNA clone encoding the 3' portion of the β ADH subunit and used it as a hybridization probe to select genomic clones from a library of human genomic DNA (21). The nucleotide sequence of all the exons contained in the human β ADH gene was determined (21) and it was observed that the gene is divided by eight introns which interrupt the amino acid coding region at codons 5/6, 39/40, 86, 115, 188/189, 275/276, 321, and 367 out of a total of 374 codons (Fig. 1).

Amino Acid Sequence Homology Among Various Alcohol Dehydrogenases.

The amino acid sequence of human ADH, predicted from the DNA sequence (21), is shown aligned with the amino acid sequences of ADHs from four other species and sheep sorbitol dehydrogenase (SDH) (Fig. 1). This comparison reveals that human ADH shares 50% residue identity with maize ADH, 25% with yeast ADH, 24% with sheep SDH, and 32% with the amino-terminal 54 residues of *Bacillus* ADH. In addition, 42% of the yeast ADH residues, 39% of the sheep SDH residues, and 70% of the *Bacillus* ADH residues (i.e. the amino-terminal 54 residues) have a match with at least one of the other structures. Thus, these enzymes could have all derived from a common ancestor that developed prior to the prokaryotic-eukaryotic branchpoint in evolution some 10^9 years ago.

Comparison of the Intron/Exon Structures of Human and Maize ADH.

If, as suggested above, the ADH genes of many organisms are derived from a common ancestral gene, then it is of interest to compare the intron/exon structures of ADH genes. As can be seen (Fig. 1), there are no residues at which both maize and human ADH share an intron. However, this may not be surprising given the length of time involved since divergence of plants and animals. The observed patterns may be explained by intron deletion, insertion, and/or mutational shifting of intron locations from an originally common site after divergence of human and maize ADH genes. Introns may have given rise to hot spots of mutational activity at bordering exon sequences and this could account for "migration" of intron positions in either a 5' or 3' direction. Thus, the human and maize ADH genes examined here may bear introns which lie in roughly homologous positions, and this may be addressed by examination of where the introns fall in relation to the secondary structural units of the enzymes.

Because of the high degree of amino acid sequence homology between horse and human ADH, the three dimensional structure of the horse enzyme (7) is probably a good approximation of the structure for human ADH. The predicted positions of β strands, α -helices, subunit interaction, zinc-binding ligands, and the coenzyme binding domain for human ADH are shown (Fig. 2) along with the positions of the introns. A similar analysis of the maize ADH gene was previously performed by Dennis et al. (9). The four cysteine

letter code) is that of horse and human ADH. The boxes surround residues which have identity with residues in at least one of the other sequences. When two different amino acids are conserved at a position, all are boxed together. The two junctions (7) of the coenzyme-binding domain with the catalytic domain are indicated. The sites of introns in the human ADH gene (21) (boxes) and maize ADH gene (9) (circles) are depicted above and below the sequence comparison, respectively.

ligands for the non-catalytic zinc atom are bounded closely by human ADH introns 3 and 4, but not by maize ADH introns. Both human and maize ADH have introns closely bordering the major continuous subunit interaction region; i.e. human ADH introns 6 and 7, and maize ADH introns 8 and 9. But, probably the most significant finding from this analysis is that introns 5 and 7 of human ADH lie very close to the coenzyme-binding domain junctions (Fig. 2), as do the maize ADH introns 4 and 9 recognized previously (8,9). Thus, in both cases introns separate the catalytic and coenzyme-binding domains. An additional human ADH intron within the coenzyme-binding domain (intron 6) falls between β strands, as do introns 5, 6, 7, and 8 within the coenzyme-binding domain of the maize ADH gene (9). Thus, the positions of human ADH introns 5, 6, and 7 support the conclusions of Brändén et al. (8) which suggest that the DNA encoding the maize ADH coenzyme-binding domain may have evolved by the fusion of DNA segments containing small exon sequences each encoding at least one intact β strand.

Intron/Exon Structures of the Nucleotide-binding Domains in Several Enzymes.

Analysis of the evolution of the nucleotide-binding domain can be further extended

Figure 2 - Relationship between the intron locations and functional domains of human ADH. The functional domains and secondary structural elements of horse ADH (7) are visualized by bars similar to those used by Dennis et al. (9) for the maize ADH gene. Distorted helices or sheets and reverse turns are not included except for the 3^{10} S helix formed by residues 305-311 (7). The numbering refers to both horse and human amino acid residues. It is assumed that human ADH has approximately the same three-dimensional structure as horse ADH. Vertical lines indicate the location of the eight human ADH introns (21).

Figure 3 - Intron/exon structures of NAD- and ATP-binding domains. The bars depict the NAD-binding domains of various dehydrogenases as well as the ATP-binding domain of pyruvate kinase; the NAD-binding domain can be subdivided into two mononucleotide-binding folds, each possessing three β strands. The solid boxes shown in the bars indicate the amino acid residues of these enzymes which comprise the β strands (β A- β F) located in the NAD- or ATP-binding domains; the positions were predicted by alignment of their amino acid sequences with those of either horse liver ADH (7), lobster GAPDH (29), dogfish LDH (29), or cat muscle PK (30) (all of which have known three-dimensional structures). The predicted positions of the β A and β B strands of *Drosophila* ADH differ from those shown previously (23), and correspond to those of Jörnvall et al. (11). The three β strands of PK are located in a region homologous to the adenine mononucleotide-binding fold of dehydrogenases (29). The amino acid codons interrupted by introns (arrows) are those determined from the DNA sequences of the genes encoding human β ADH (21), maize ADH 1 (9) and ADH 2 (22) (identical locations), *Drosophila* ADH (23,28), chicken GAPDH (24), mouse LDH-A (26), and chicken PK (27). The cross-hatched bars indicate 5'- or 3'-untranslated (UT) regions. The intron in the 5'-UT region of the *Drosophila* ADH gene lies 36 bp upstream of the AUG translation start codon (28), and that of the GAPDH gene lies 18 bp upstream (24). The position of the 5'-UT intron in LDH has not been determined, but is assumed to lie 24 bp upstream of the start codon where a DNA sequence matching that of the consensus splice acceptor is located (26). Gaps of about 20 amino acids were introduced into the GAPDH and LDH structures to facilitate alignment of the β strands; they correspond to loops not involved in this comparison.

by comparison of the intron/exon structures of human and maize ADH with those of Drosophila ADH, chicken GAPDH, mouse LDH, and chicken PK (Fig. 3). The NAD-binding domains of GAPDH and LDH resemble very closely that of ADH, while the ATP-binding domain of PK is homologous to the adenine mononucleotide binding fold of the NAD-binding domain (29). Interestingly, all six genes possess introns that demarcate the borders of the NAD-binding domain (the ATP-binding domain for PK), and collectively there are examples of introns separating all six conserved β strands; i.e. maize ADH and chicken GAPDH have all but the β E and β F strands separated by introns, and mouse LDH has an intron separating these two β strands (Fig. 3). In no case is a β strand divided, although there are a few introns lying exactly at one end of a β strand, i.e. Drosophila ADH residue 32 as well as GAPDH residues 7 and 145. It can be seen (Fig. 3) that Drosophila ADH, chicken GAPDH, and mouse LDH have NAD-binding domains positioned in the N-terminal half of the subunit, whereas the other enzymes have this domain closer to the C-terminus. To explain this one might envision rearrangements which juxtaposed a block of exons encoding the coenzyme-binding domain to a position upstream of a coding region for a catalytic domain. The fact that these three genes possess introns in their 5'-untranslated regions (just upstream from the coenzyme-binding domain) supports this possibility. The 5'-untranslated introns of these genes may thus represent the vestiges of exon-shuffling of the nucleotide-binding domain to an N-terminal location. These introns may have been retained because of some regulatory significance which they imparted on the gene.

DISCUSSION

The locations of introns in the human ADH gene as well as genes for related enzymes argues for the intron-dependent evolution of their nucleotide-binding domains. It appears quite possible that a single primeval nucleotide-binding domain could have evolved at an early stage, and then become attached via exon-shuffling to various catalytic domains giving rise to dehydrogenases and kinases with differing substrate specificities. This includes the gene of short-chain Drosophila ADH (23,28) where intron patterns similar to those in genes of long-chain ADHs (9,21,22) are now recognized. The results further provide a firm link between the evolution of the NAD-binding domain and the ATP-binding domain, thus supporting the conclusions of others (27,29). The intron/exon structures discussed above are consistent with the following hypotheses: (a) that wide relationships exist among enzymes having a mononucleotide binding unit (29); (b) that introns were present prior to the prokaryotic-eukaryotic branchpoint in evolution (2); and (c) that introns accelerated protein evolution via exon-shuffling (1). The recent identification of an intron in the thymidylate synthase gene of bacteriophage T4 (31) further suggests that introns are evolutionarily old.

The data presented here does not preclude an intron insertion model for the generation of the intron/exon structures observed for nucleotide-binding domains. If insertions were slightly mutagenic, possibly due to duplications of several base pairs at the site of insertion as is commonly observed for transposable elements, introns in regions encoding important functional or structural elements would be selectively lost. Thus, the intron/exon structures observed could be due to random insertion and non-random loss. In such a model the β -strands in the nucleotide-binding domains could have deleted any introns which may have inserted into them, leaving a random assortment of introns between the β -strands in regions which are of little importance structurally or functionally. However, we argue that a model which proposes that introns were present from the beginning of gene evolution makes sense from a logistical viewpoint since this reduces the number of random events which were required to produce the similar patterns of introns we observe today in genes encoding nucleotide-binding domains. Each individual species would thus have a pattern of introns which primarily reflects intron migration or deletion from the original pattern with insertions accounting for some but not all introns. This point is particularly driven home by the similarity of the intron/exon structures of the NAD-binding domains in maize ADH and chicken GAPDH which both have introns separating all the β strands except the β E and β F strands (Fig.3). Following this model it appears that the intron/exon structures in the coenzyme-binding domains of the maize ADH gene and chicken GAPDH gene more closely resemble that of the ancestral gene, with the human ADH gene representing a streamlined form that has lost most of the introns which separate the β strands (Fig. 3). However, there must have been considerable migration of the remaining introns if the model is to be correct. In our opinion, the similarity of intron locations among these genes is difficult to explain solely on the basis of random insertions throughout time and is better explained by the intron-dependent evolution of a commonly shared nucleotide-binding domain.

A model for the intron-dependent evolution of NAD- and ATP-binding domains is presented (Fig. 4). Based on tertiary structure comparisons Rossman et al. (29) proposed a common evolutionary origin for these domains and suggested they arose during precellular evolution. Since there are examples of introns separating each of the β strands (see Fig. 3), it is possible that an ancestral nucleotide-binding domain evolved by duplication of a single primordial exon (or recombination of distinct exons) encoding a single β strand (Fig. 4). As pointed out by Lonberg and Gilbert (27), there is no evidence other than secondary structure homology to suggest duplication, but mutations during the long period of time involved could have eroded the primary sequence homology. As suggested by the exon-shuffling hypothesis (1), the joining of exons was an imprecise process, leaving long stretches of non-informational sequence (introns). Genes possessing exons encoding NAD- and ATP-binding domains may have joined to co-evolving genes possess-

Figure 4 - Proposed evolutionary history of the NAD- and ATP-binding domains in dehydrogenases and kinases. This scheme is based upon our results and those of others (8,27,29). Concerning the higher eukaryotic genes indicated here, only the exons comprising the NAD- or ATP-binding domains were counted (see Fig. 3). The genes encoding yeast ADH (17-19), GAPDH (32), and PK (33) do not contain introns. The bacterial counterparts for these genes have not been analyzed but presumably lack introns.

ing exons encoding catalytic domains at an early stage in evolution, in agreement with previous suggestions (29), giving rise to genes for various dehydrogenases and kinases (Fig. 4). Prokaryotes and lower eukaryotes may have eliminated most of their introns (via random deletions or reverse transcription of mRNA followed by gene conversion) due to the selective advantages of streamlining their genomes. Perhaps higher eukaryotes found it selectively advantageous to retain many introns especially those which fall at each end of the NAD- and ATP-binding domains, thus preserving these domains as potentially mobile units discrete from their neighboring catalytic domains. Such conserved introns may have ensured that evolution via exon-shuffling could continually occur to produce new dehydrogenases and kinases, thus contributing to the variability which allowed complex organisms to develop.

ACKNOWLEDGMENTS

We are indebted to Steven S.-L. Li for communicating the mouse lactate dehydrogenase-A intron/exon structure prior to publication. G.D. was supported by an NIH postdoctoral fellowship (GM-08733) and a research grant from the National Institute of

Alcohol Abuse and Alcoholism (AA-06622). G.W.H. was supported by NIH grant GM-32112 and H.J. by a research grant from the Swedish Medical Research Council (project 13X-3532).

§To whom correspondence should be addressed

REFERENCES

1. Gilbert, W. (1978) *Nature* 271, 501.
2. Doolittle, W. F. (1978) *Nature* 272, 581-582.
3. Darnell, J. E. (1978) *Science* 202, 1257-1260.
4. Cavalier-Smith, T. (1978) *J. Cell. Sci.* 34, 247-278.
5. Crick, F. (1979) *Science* 204, 264-271.
6. Sharp, P. A. (1985) *Cell* 42, 397-400.
7. Eklund, H., Nordström, B., Zeppezauer, E., Söderlund, G., Ohlsson, I., Boiwe, T., Söderberg, B.-O., Tapia, O., Brändén, C.-I., and Åkeson, Å. (1976) *J. Mol. Biol.* 102, 27-59.
8. Brändén, C.-I., Eklund, H., Cambillau, C., and Pryor, A. J. (1984) *EMBO J.* 3, 1307-1310.
9. Dennis, E. S., Gerlach, W. L., Pryor, A. J., Bennetzen, J. L., Inglis, A., Llewellyn, D., Sachs, M. M., Ferl, R. J., and Peacock, W. J. (1984) *Nucleic Acids Res.* 12, 3983-4000.
10. Brändén, C.-I., Jörnvall, H., Eklund, H., and Furugren, B. (1975) *The Enzymes*, ed. Boyer, P. D. (Academic, New York), Vol. 11, pp. 103-190.
11. Jörnvall, H., von Bahr-Lindström, H., Jany, K.-D., Ulmer, W., and Froeschle, M. (1984) *FEBS Lett.* 165, 190-196.
12. Eklund, H., Horjales, E., Jörnvall, H., Brändén, C.-I., and Jeffery, J. (1985) *Biochemistry*, in press.
13. Jeck, R., Woenckhaus, C., Harris, J. L., and Runswick, M. J. (1979) *Eur. J. Biochem.* 93, 57-64.
14. Jörnvall, H. (1970) *Eur. J. Biochem.* 16, 25-40.
15. Hempel, J., Bühler, R., Kaiser, R., Holmquist, B., de Zalski, C., von Wartburg, J.-P., Vallee, B., and Jörnvall, H. (1984) *Eur. J. Biochem.* 145, 437-445.
16. Hempel, J., Holmquist, B., Fleetwood, L., Kaiser, R., Barros-Soderling, J., Bühler, R., Vallee, B., and Jörnvall, H. (1985) *Biochemistry*, 5303-5307.
17. Bennetzen, J. L., and Hall, B. D. (1982) *J. Biol. Chem.* 257, 3018-3025.
18. Russell, D. W., Smith, M., Williamson, V. M., and Young, E. T. (1983) *J. Biol. Chem.* 258, 2674-2682.
19. Russell, P. R. and Hall, B. D. (1983) *J. Biol. Chem.* 258, 143-149.
20. Jörnvall, H., von Bahr-Lindström, H., and Jeffery, J. (1984) *Eur. J. Biochem.* 140, 17-23.
21. Duester, G., Smith, M., Bilanchone, V., and Hatfield, G. W. (1986) *J. Biol. Chem.*, in press.
22. Dennis, E. S., Sachs, M. M., Gerlach, W. L., Finnegan, E. J., and Peacock, W. J. (1985) *Nucleic Acids Res.* 13, 727-743.
23. Benyajati, C., Place, A. R., Powers, D. A., and Sofer, W. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 2717-2721.
24. Stone, E. M., Rothblum, K. N., Alevy, M. C., Kuo, T. M., and Schwartz, R. J. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 1628-1632.

25. Stone, E. M., Rothblum, K. N., and Schwartz, R. J. (1985) *Nature* 313, 498-500.
26. Li, S. S.-L., Tiano, H. F., Fukasawa, K. M., Yagi, K., Shimizu, M., Sharief, F. S., Nakashima, Y., and Pan, Y. E. (1985) *Eur. J. Biochem.* 149, 215-225
27. Lonberg, N. and Gilbert, W. (1985) *Cell* 40, 81-90.
28. Benyajati, C., Spoerel, N., Haymerle, H., and Ashburner, M. (1983) *Cell* 33, 125-133.
29. Rossman, M., Liljas, A., Brändén C.-I., and Banaszak, L. (1975) *The Enzymes*, ed. Boyer, P. D. (Academic, New York), Vol. 11, pp. 61-102.
30. Stuart, D. C., Levine, M., Muirhead, H., and Stammers, D. K. (1979) *J. Mol. Biol.* 134, 109-142.
31. Chu, F. K., Maley, F. G., Maley, F., and Belfort, M. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81, 3049-3053.
32. Holland, J., and Holland, M. (1980) *J. Biol. Chem.* 255, 2596-2605.
33. Burke, R. L., Tekamp-Olson, P., and Najarian, R. (1983) *J. Biol. Chem.* 258, 2193-2201.