J Mol Evol (1991) 33:194–203

Journal of Molecular Evolution

Drosophila Fat Body Protein P6 and Alcohol Dehydrogenase are Derived from a Common Ancestral Protein

Luce Rat,[1] Michel Veuille,[2] and Jean-Antoine Lepesant[1]

[1] Institut Jacques Monod, CNRS et Université Paris 7, 75251 Paris-cedex 05, France
[2] Laboratoire de Biologie et Génétique Evolutives, CNRS, 91198 Gif-sur-Yvette, France

Summary. *Drosophila melanogaster* alcohol dehydrogenase is an example of convergent evolution: it is not related to the ADHs of other organisms, but to short-chain dehydrogenases, which until now have been found only in bacteria and in mammalian steroid hormone metabolism. We present evidence that the *Drosophila* ADH is phylogenetically more closely related to P6, another highly expressed protein from the fat body of *Drosophila,* than it is to the short-chain dehydrogenases. The polypeptide sequence of P6 was inferred from DNA sequence analysis. Both ADH and P6 polypeptides have retained a high structural similarity with respect to the Chou–Fasman prediction of secondary structure and hydropathy. P6 is also homologous to the 25-kd protein from the fat body of *Sarcophaga peregrina,* whose sequence we have reexamined. The evolution of the P6-ADH family of proteins is characterized by a dramatic increase in the methionine content of P6. Methionine accounts for 20% of P6 amino acids. This is in contrast with the absence of this amino acid in mature ADH. There is evidence that P6 and the 25-kd protein have undergone a parallel and independent enrichment in methionine. When corrected for this, the rate of amino acid replacement shows that the P6–25-kd lineage diverged from insect ADH shortly before the divergence of the ADH gene (Adh) from its 3′-duplication (Adh-dup).

Key words: Short-chain dehydrogenases — Molecular evolution

Offprint requests to: J.-A. Lepesant

Abbreviations: Adh, alcohol dehydrogenase gene; ADH, alcohol dehydrogenase polypeptide; Adh-dup, 3′ duplication of *Drosophila melanogaster* Adh gene; ADH-DUP, polypeptide sequence inferred from the Adh-dup sequence

Introduction

Drosophila alcohol dehydrogenase (ADH; alcohol: NAD^+ oxidoreductase, EC 1.1.1.1) poses an interesting evolutionary problem. This enzyme has become an important focus for molecular biology and molecular evolution studies (Kreitman 1987; Sofer and Martin 1987; Chambers 1988; Sullivan et al. 1990). Its origin however remains unknown. It is not related to the mammalian or plant ADHs, which are a close group of highly homologous proteins (Yokoyama et al. 1990). It differs from them in needing no zinc and in having a shorter polypeptide sequence. *Drosophila* ADH has been characterized as a short-chain dehydrogenase (Jörnvall et al. 1981, 1984). This group of enzymes includes a heterogeneous set of proteins found in bacteria and mammals. This group is composed of glucose dehydrogenase from *Bacillus megaterium* (Jany et al. 1984), ribitol dehydrogenase from *Klebsiella aerogenes* (Jörnvall et al. 1981), glucitol-6-phosphate dehydrogenase from *Escherichia coli* (Yamada and Saier 1987), dihydrodiol dehydrogenase from *Pseudomonas pseudoalcaligenes* (Furukawa et al. 1987), 20-β-hydroxysteroid dehydrogenase from *Streptomyces hydrogenans* (Marekov et al. 1990), 17-β-hydroxysteroid dehydrogenase, 11-β-corticosteroid dehydrogenase (Baker 1990), and 15-hydroxyprostaglandin dehydrogenase (Krook et al. 1990) from mammals, and several other proteins from either mammals (Navre and Ringold 1988) or bacteria (Debellé and Sharma 1986; Coleman et al. 1988; Hallam et al. 1988) of unknown metabolic function. These proteins, however, are too distantly related to insect ADH to provide any information on how and when it acquired its catalytic activity. We present a new protein from the *Drosophila* fat body, P6, which is a very close relative of *Drosophila* ADH.

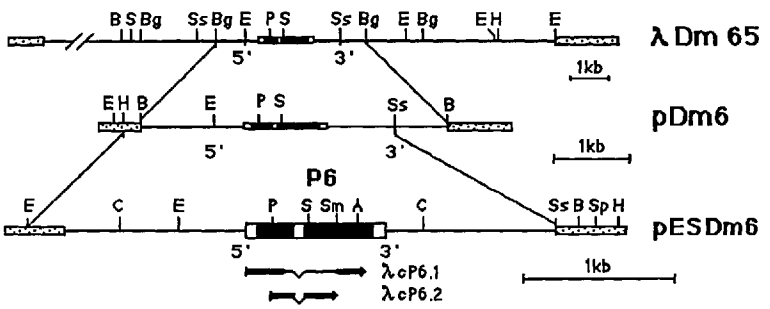

Fig. 1. Molecular organization of the P6 gene and flanking sequences. Simplified restriction map of the *Drosophila* DNA insert contained in the recombinant bacteriophage λ-Dm 65 (Lepesant et al. 1982) and in the pDm6 plasmid (B, Bam HI; S, Sal I; Bg, Bgl II; Ss, Sst I; E, Eco RI; P, Pst I; H, Hind III; Sm, Sma I; C, Cla I; Sp, Sph I). A 3.5-kb Hind III–Sst I fragment of pDm6 was subcloned into the pEMBL 18 (+) vector in one (pESDm6) or the opposite orientation (pSSDm6, not shown). Stippled boxes indicate vector sequences. The open and filled boxes indicate the untranslated and translated coding sequences, respectively. Complementary DNA inserts contained in the recombinant bacteriophages λ-cP6-1 and λ-cP6-2 are indicated by arrows, the thickened portion of them indicating the sequenced regions.

The *Drosophila melanogaster* P6 protein has been characterized as a major fat body-specific protein of the third larval instar (Lepesant et al. 1982; Deutsch et al. 1989). P6 and ADH were not previously thought to be related. Even though ADH and P6 have a very similar molecular weight, P6 is methionine-rich, whereas this amino acid is virtually lacking in ADH (Benyajati et al. 1981). Only when we sequenced the gene coding for P6, did it become apparent that the two proteins share a strong sequence similarity. This indicates that the P6, Adh, and the related Adh-3′ duplication (Adh-dup) (Schaeffer and Aquadro 1987) genes have evolved from a common ancestor that may represent an intermediate in the emergence of insect ADH.

Materials and Methods

Unless otherwise noted, nucleic acids were handled according to standard protocols (Sambrook et al. 1989).

The clones used in this study were isolated from a genomic library of randomly cleaved *D. melanogaster* DNA fragments inserted into the λ vector Charon 4 (Maniatis et al. 1978); these clones were identified by their hybridization to a probe of poly(A)$^+$ RNA from third-instar larval fat bodies (Lepesant et al. 1982).

Sequence Determination. The dideoxy chain termination method of Sanger et al. (1977) was used to determine the sequence of deletion derivatives of the pSSDm6 and pESDm6 plasmids (Fig. 1) generated according to Henikoff (1984), and the sequence of the complementary DNA (cDNA) inserts contained in the λ-cP6.1 and λ-cP6.2 bacteriophages. Sequence data were assembled and analyzed using the programs available at the C.I.T.I.2 (Paris).

Mapping of the 5′ and 3′ Ends of the P6 Transcripts. Total cellular RNA was isolated from fat bodies of late third-instar larvae of the Canton-S stock (Lepesant et al. 1982). Primer extension was carried out essentially according to Calzone et al. (1987) using the [^{32}P]5′-end-labeled primer (5′TCTTACCGGT CCAGTCAAAC3′) and M-MLV reverse transcriptase. For S1 nuclease mapping of the 5′ end, the labeled probe (Fig. 3) and total RNA were coprecipitated with ethanol, dried under vacuum, dissolved in 20 μl of hybridization buffer (40 mM PIPES pH 6.4, 400 mM NaCl, 1 mM EDTA, 70% formamide), and incubated for 3 min at 80°C and then 3 h at 44°C. Annealed samples were diluted 10-fold in ice-cold nuclease buffer (50 mM sodium acetate, 280 mM NaCl, 4.5 mM $ZnSO_4$, and 20 mg/ml salmon sperm DNA) and further incubated at 37°C for 30 min in the presence of 1800 units/ml of S1 nuclease. After phenol–chloroform extractions and ethanol precipitation, the protected DNA fragments were analyzed by electrophoresis in 6% denaturing polyacrylamide gels. For S1 nuclease mapping of the 3′ end the same procedure was applied except that the denaturation and annealing temperatures were 75°C and 45°C, respectively, and the S1 nuclease concentration was 2200 units/ml.

Results

Structure of the P6 Gene

The structure of the P6 gene was determined by sequencing a 2796-bp genomic DNA fragment encompassing the complete coding sequence (Figs. 1 and 2). Primer extension, S1 nuclease protection (Fig. 3, panel 5′ end), and cDNA sequencing indicated the presence of three potential closely spaced mRNA start sites (Fig. 2). The 3′ end of the P6 transcript was mapped by S1 protection to position +975 (Fig. 3, panel 3′ end). Two overlapping variants of the canonical AATAAA polyadenylation site upstream and a 14-bp cluster of T and G nucleotides immediately downstream conform to structural features around mRNA polyadenylation sites as proposed by Birnstiel et al. (1985). Sequencing of cDNAs revealed the presence of a 60-bp intron that interrupts the coding sequence. Flanking sequences upstream from the P6 mRNA start site were compared with those of other genes expressed in the larval fat body, P1 (Maschat et al. 1990), LSP-1-α-, β-, γ- (Delaney et al. 1986), LSP-2 (Mousseron, personal communication), and Adh (Benyajati et al. 1981). Two AT-rich regions (Fig. 4A and B) exhibit partial identity with sequences upstream of the P1 gene; a third 28-bp region (Fig. 4C) located imme-

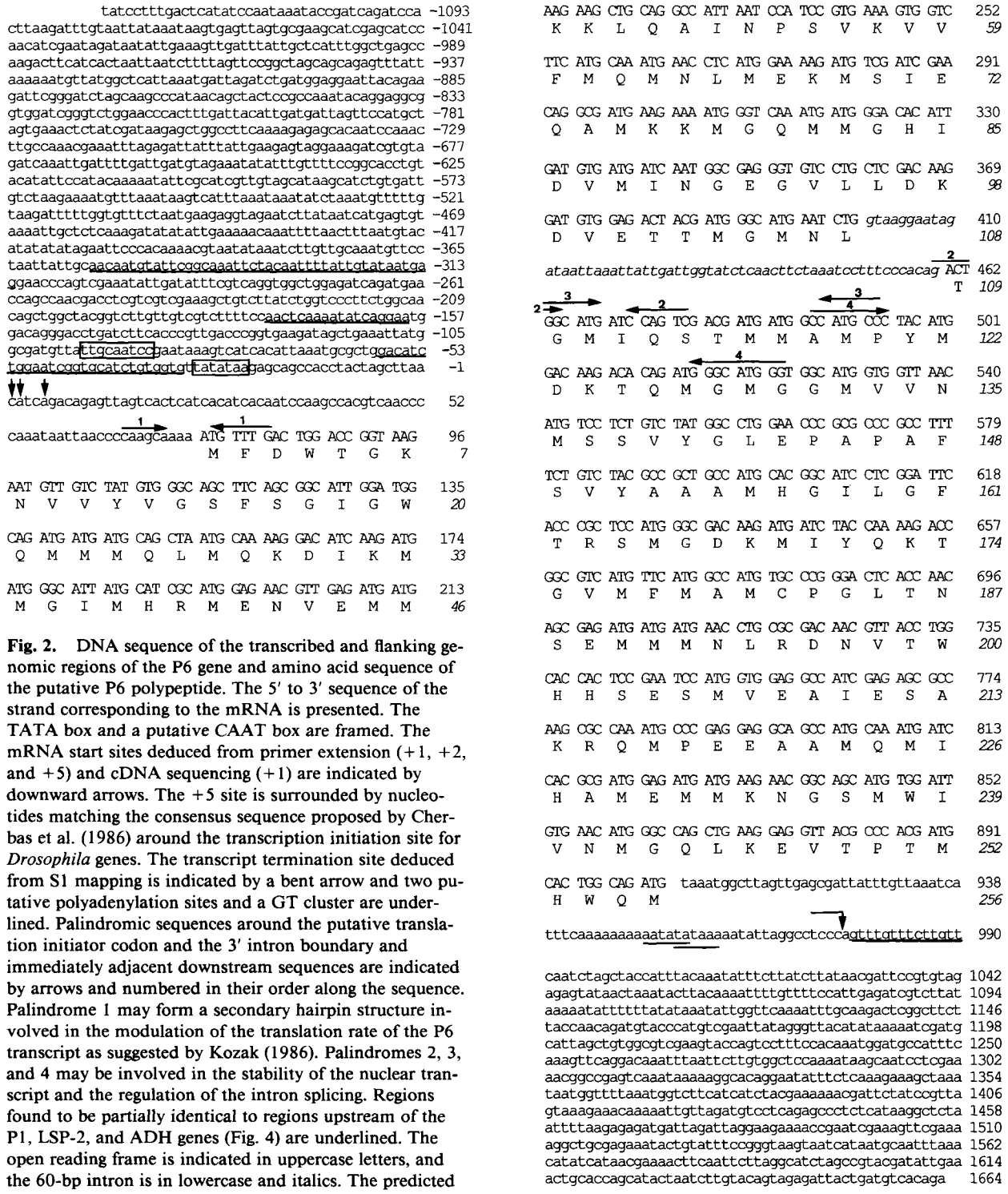

Fig. 2. DNA sequence of the transcribed and flanking genomic regions of the P6 gene and amino acid sequence of the putative P6 polypeptide. The 5′ to 3′ sequence of the strand corresponding to the mRNA is presented. The TATA box and a putative CAAT box are framed. The mRNA start sites deduced from primer extension (+1, +2, and +5) and cDNA sequencing (+1) are indicated by downward arrows. The +5 site is surrounded by nucleotides matching the consensus sequence proposed by Cherbas et al. (1986) around the transcription initiation site for *Drosophila* genes. The transcript termination site deduced from S1 mapping is indicated by a bent arrow and two putative polyadenylation sites and a GT cluster are underlined. Palindromic sequences around the putative translation initiator codon and the 3′ intron boundary and immediately adjacent downstream sequences are indicated by arrows and numbered in their order along the sequence. Palindrome 1 may form a secondary hairpin structure involved in the modulation of the translation rate of the P6 transcript as suggested by Kozak (1986). Palindromes 2, 3, and 4 may be involved in the stability of the nuclear transcript and the regulation of the intron splicing. Regions found to be partially identical to regions upstream of the P1, LSP-2, and ADH genes (Fig. 4) are underlined. The open reading frame is indicated in uppercase letters, and the 60-bp intron is in lowercase and italics. The predicted amino acid sequence of the P6 polypeptide deduced from the open reading frame is indicated (upper case) under the nucleotide sequence.

diately upstream of the P6 TATA box shows sequence identity with sequences at various locations upstream of the mRNA start site of the P1, Adh, and LSP-2 genes. No significant sequence similarity with the LSP-1 gene was observed. The functional significance of these sequence identities remains to be established, but it is interesting to note that the Adh sequence thus detected overlaps an essential regulatory element, P_0, which was first detected by Heberlein et al. (1985) and shown by Shen et al. (1989) to have enhancer properties.

Homologous Proteins

The nucleotide sequence of P6 predicts a 256-amino acid polypeptide (29 kd). A systematic search for

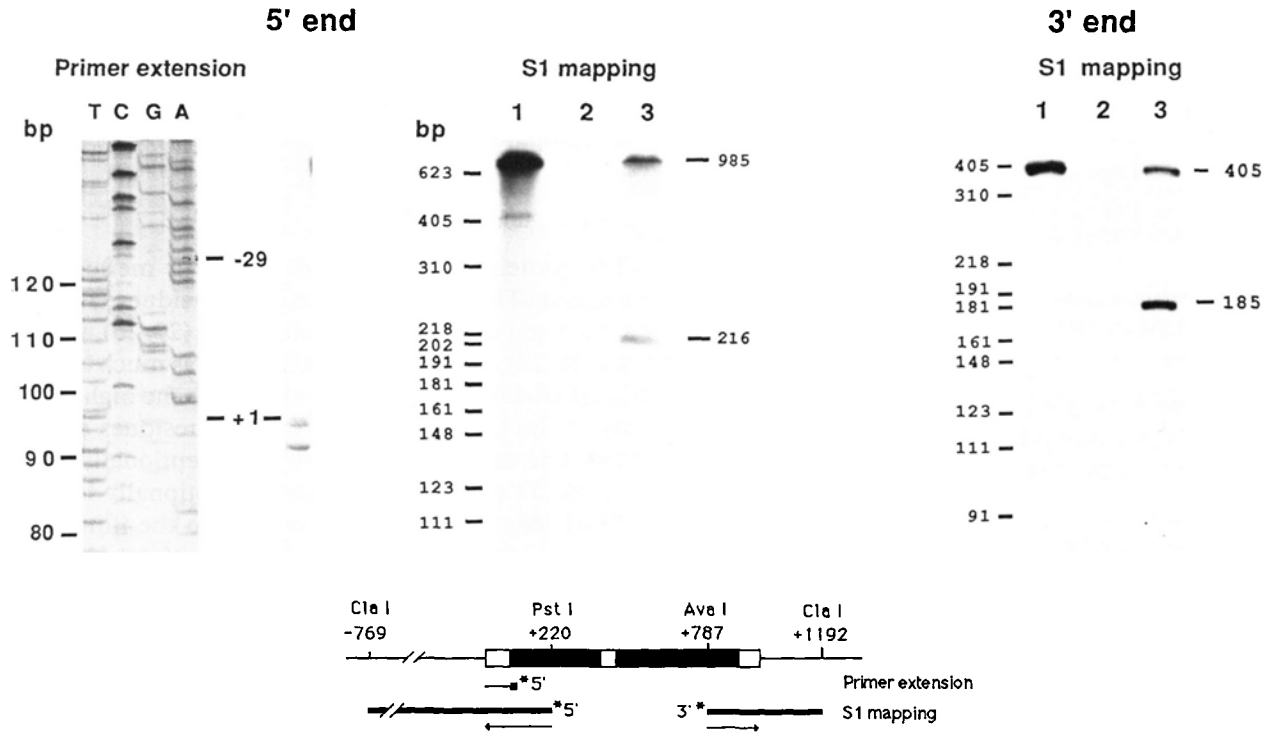

Fig. 3. Mapping of the 5′ and 3′ ends of the P6 transcript. *5′ end:* Primer extension was carried out as described in the Materials and Methods using 67 fmoles (8 × 10^4 cpm) of labeled primer and 16 μg of total RNA from late third-instar fat body. Sanger sequencing reactions performed with the same primer were run in parallel as a size standard. A +1 indicates the position of the most likely upstream transcription start site. A −29 indicates the position of the TATA box. S1 mapping was carried out using 30 fmoles (2.5 × 10^4 cpm) of the 985-bp 5′ end-labeled Pst I–Cla I fragment as a probe, 10 μg of total RNA and 1800 units/ml of S1 nuclease. Lanes 1: labeled probe; 2, control with total embryonic RNA; 3: probe and protected fragment. *3′ end:* S1 mapping was carried out using 8 fmoles (4.5 × 10^4 cpm) of the 405-bp Ava I–Cla I 3′ end-labeled fragment as a probe, 5 μg of total RNA and 2200 units/ml of S1 nuclease. Lanes 1, labeled probe; 2: control with embryonic RNA; 3, probe and protected fragment. The diagram indicates the position and extent of the probes and protected or extended fragments on a partial restriction map of the P6 locus. Numbers refer to the positions of the restriction endonuclease cleavage sites relative to the start of transcription. The transcribed region is indicated by open and filled boxes. The untranslated region and the 60-bp intervening sequence are indicated by open boxes.

similar protein sequences led us to compare P6 to the following three proteins. A consensus alignment (using the Kanehisa program, available at the C.I.T.I.2, Paris) is shown in Fig. 5:

1) A strong similarity (66% amino acids) exists between P6 and the 25-kd protein (25-kDa) from *S. peregrina,* a major fat body protein of unknown function, the sequence of which was predicted from its DNA sequence by Matsumoto et al. (1985). The similarity extends further than the polypeptide inferred from the published DNA sequence and it also encompasses an open reading frame (ORF) immediately 3′ to the position formerly considered to be a stop codon (indicated in Fig. 5 by a star). It continues over 222 more base pairs, and predicts a 258-amino acid polypeptide, which is about the same length as P6. This finding seems to solve the puzzling fact that the known sequence of 25-kd appeared to be 30% shorter than expected (Matsumoto et al. 1985; Henikoff and Wallace 1988). This, along with the nearly identical hydropathy of the two proteins on the whole of P6 amino acids (Fig. 6), led us to conclude that the stop codon formerly found at position 185 resulted from a frameshift due to a single base pair deletion in the sequenced clone (Fig. 5, star).

2) There is a lower but significant similarity (28% amino acids) between P6 and *Drosophila* ADH (Benyajati et al. 1981), and with the polypeptide (30% amino acids) inferred from Adh 3′-duplication (3′-ORF, or Adh-dup), which will be called hereafter ADH-DUP (Schaeffer and Aquadro 1987). No published sequence exists for *D. melanogaster* Adh-dup; the sequences hereafter referred to as Adh-dup and Adh are from *Drosophila mauritiana* (Cohn and Moore 1988), a very closely related species. All four proteins, P6, 25-kd, ADH-DUP, and ADH are of similar length and differ only by a few gaps. The intron positions of their genes match exactly (Fig. 7): 25-kd has three introns, two of which (introns 1

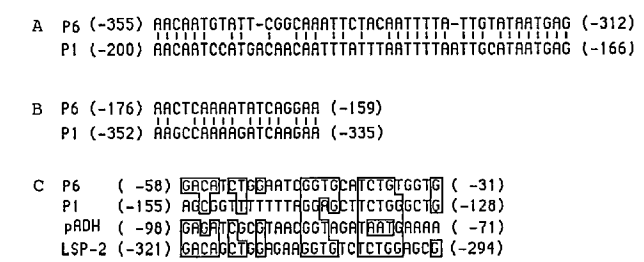

Fig. 4. Alignment of the P6 flanking sequences with other genes expressed in the fat body

and 3) are shared by ADH-DUP (Schaeffer and Aquadro 1987) and ADH (Cohn and Moore 1988). The third (intron 2) is shared by P6.

Protein Sequence Conservation

When compared with 25-kd, the middle third of P6 (residues 90–179) is more highly conserved than the other two-thirds, with the C-terminal portion being slightly more variable ($\chi^2 = 16.05$, $P < 0.001$, df = 2). The same trend is observed when comparing P6 to either ADH-DUP or ADH. A similar pattern has been found for the ADHs of *D. melanogaster* and *Drosophila pseudoobscura* (Schaeffer and Aquadro 1987), where both amino acid replacements and silent nucleotide changes are more frequent in the third exon and less frequent in the second exon. In addition, the hydropathy profile (Hopp and Woods 1981) of all four proteins (Fig. 6), which is nearly identical in P6 and 25-kd, is highly conserved in its middle part with regard to ADH-DUP and ADH.

Secondary Structure Conservation

Structural homologies are further shown by the β sheet/α helix protein organization. The coenzyme-binding domain of *Drosophila* ADH is known to lie within the first 140 amino acids (Benyajati et al. 1981). This region comprises six β-strands, which are conventionally numbered A through F. Strands A through C belong to the adenine-binding domain, whereas strands D through F belong to the nicotinamide-binding domain. A Chou–Fasman prediction of the secondary structure (Chou and Fasman 1978) of P6 showed that the β-strands lie in positions similar to those of Adh (Fig. 5, upper line).

Codon Usage

The codon usage of *Drosophila* Adh is known to be highly biased (Shields et al. 1988). This same bias is observed for P6. Differences in the usage of A-, C-, G-, and T-ending codons (excluding those for methionine and tryptophan) are nonsignificant between P6 and Adh ($\chi^2 = 6.89$, df = 3), and are highly significant between either P6 or Adh and Adh-dup (P6: $\chi^2 = 23.04$, $P < 0.001$; Adh: $\chi^2 = 54.2$, $P < 0.001$; df = 3). This is due to a bias toward C- and G-ending codons in P6 (75.0%) and in Adh (81.2%), as compared to Adh-dup (54.9%).

Amino Acid Content

The proteins differ in acidity and in methionine content. The ratio of acidic/basic residues (Asp + Glu/Arg + Lys) is about unity in P6 (23/20), in 25-kd (21/22), and in ADH (22/23), and is much higher in ADH-DUP (31/21), mainly due to the high acidity of the C-terminal part (6 acidic residues out of 16). The methionine content is exceptionally high in P6 (51 residues, 20%) and exceptionally low in ADH (one residue, corresponding to the initiation codon), with intermediate values in 25-kd (29 residues, 11.3%) and in ADH-DUP (13 residues, 4.8%). Although P6, 25-kd, and ADH-DUP all seem to be methionine-rich, their methionine positions are poorly conserved. A large proportion of substitutions between P6 and 25-kd (46 out of 112) are due to changes at methionine positions in either direction.

Discussion

P6 and Adh are among the predominant proteins in the *Drosophila* fat body, each amounting to around 1% of the soluble protein content of this organ at their respective peaks of expression. The reason why such large amounts of these proteins are produced is unclear. The catalytic function of ADH is known, but its physiological role is not. It has been shown to be involved in ethanol detoxification in *D. melanogaster* (David 1977). Adaptation to ethanol-rich environments is however a peculiarity of this species, and therefore a recent feature in *Drosophila* evolution (Lachaise et al. 1988), whereas the divergence of ADH from P6 is clearly, from the above data, a much earlier event in the history of insects.

Methionines and the Evolution of P6

The striking difference in the methionine content of P6 and ADH probably results from a new function acquired by P6 after its divergence from their common ancestor. Two observations that support the hypothesis that this is a recent phenomenon include (1) the difference in the methionine load of P6 and 25-kd, and (2) the mismatch of their methionine sites. The percent methionine content of these polypeptides seems to behave like a quantitative character, with no strict specificity to the location of this residue within the polypeptide. A high methionine

```
                          1         10        20        30 <intron>    40        50        60        70
                                  ( A ) adenine binding        ( B )   site        ( C )
P6     :D.melanogaster    M-FDWTGKNVVYVGSFSGIGWQMMMQLMQKDIKMM     GIMHRMENVEMMKKLQAINPSVKVVFMQMNLME-KMSIE
25kDa  :S.peregrina       .-M..NN.......G...F.Y.VCQMM.K.PM.HL <1> IVCS.......L.......T....M.V...IAD-YA..V
ADHDUP:      D.           .-..L...H.C..ADCG..ALETSKV..T.N.AKL <1> A.LQST..PQAIAQ..S.K..TQIF.WTYDVTMAREEMK
ADH    :mauritiana       .A.TL.N...IF.AGLG...LDTSKE.VKR.L.NL <1> V.LD.I..PAAIAE......K.T.T.YPYDVTVPIAETT

                              80        90        100    <intron>110       120       130       140
                                    ( D )   nicotinamide binding ( E ) site          (     F     )
P6     :D.melanogaster    QAMKKMGQMMGHIDVMINGEGVLLDKDVETTMGMNL <1> TGMIQSTMMAMPYMDKTQMGMGGMVVNMSSVYGLEPA
25kDa  :S.peregrina       KGV.QVIGHV..V..L...V.G.A.......VAV.. <2> ..L.NT.L.F........S.H.....SI........G
ADHDUP:      D.           KYFDEVMVQ.DY...L...ATLCDENNIDA.INT..      ...MNTVATVL....RKMG.S..LI..VT..I..D.S
ADH    :mauritiana       KLL.TIFAKLKTV..L...A.I.D.HQI.R.IAV.Y      ..LVNT.TAILDFW..RKG.P..IIC.IG..T.FNAI

                          150       160   <intron> 170       180       190       200           210
P6     :D.melanogaster    PAFSVYAAAMHGILGFTRSMG     DKMIYQKTGVMFMAMCPGLTNSEMMMNLRDNVTWH---HSESMV-EAIESAK
25kDa  :S.peregrina       ......S..K..GI......A <3> .EHL.H....A..CI..A*.ST.L...K..MNWMKWVP...E.W-KMVMD..
ADHDUP:      D.           .V.CA.S.SKF.VI.....LA <2> .PLY.SQN..AV..V.C.P.RVFVDRE.KA------FLEYGQSFADRLRR.P
ADH    :mauritiana       YQVP..SGTKAAVVN..S.LA <2> -.LA-PI...TAYTVN..I.RTTLVHKFNS------WLDV.PQ.A.KLLAHP

                               220       230       240       250
P6     :D.melanogaster    RQMPEEAAMQMIHAMEMMKNGSMWIVNMGQLKEVTPTMHWQM
25kDa  :S.peregrina       M.T...C.VN.MT...QA...AIY.CSTSGM..I...VYMH
ADHDUP:      D.           C.STSVCGQNIVN.I.RSE..QI..ADK.G.EL.KLHWY.H.ADQFVHYMQSNDEEDQD
ADH    :mauritiana       T.PSLAC.ENFVK.I.LNQ..AI.KLDLST.EA.QW.K..DSGI
```

Fig. 5. Alignment of the amino acid sequences of P6 and three homologous insect proteins. The sequences are derived from DNA, not from mature proteins. First line, numbered amino acid positions in P6; second line, location of P6 β-strands considered homologous to those of the *Drosophila* ADH NAD^+-binding site; parentheses, location of the introns (not including the intron located in the untranslated leader of Adh); dots, identical amino acids in P6 and the other proteins; star, position considered to be a stop codon by Matsumoto et al. (1985).

content is very unusual in proteins. In a survey of *Drosophila* proteins from GenBank release 35 (Grantham et al. 1986), methionine accounts for only 2.58% (SE = 0.16%) of the residues, a value conforming to that (2.99%) of a recent survey of Protein Sequence Database (Simon and Cserzö 1990). According to Dayhoff's (1978) computation of amino acid substitution rates, the relative rarity of methionine is due to its high replacement rate, especially toward leucine. For this reason, the increase in the methionine content of P6 and 25-kd is more likely to result from selection than from random drift.

Phylogeny of the P6–ADH Family of Proteins

Comparison of the intron/exon patterns suggests that the genes evolved from a common ancestor that possessed three introns at the positions currently seen in 25-kd. One branch, leading to Adh and Adh-dup, subsequently lost the middle intron, whereas the other leading to P6 lost the other two introns. This suggests that P6 results from an earlier duplication of a gene ancestral to Adh and Adh-dup. The proteins could therefore be grouped according to the following scheme: [P6, 25K, (ADH, ADH-DUP)].

The rate of this evolution can be assessed using the unweighted pair-group method using arithmetic averages (UPGMA). This method yields a tree reflecting the actual evolutionary lineage only if the average rate of substitution is constant over the branches of the tree. Whether this holds true for strictly homologous proteins is still a matter of debate. In the present example, the "homology" category refers to two different kinds of relationships between proteins. Two ADHs from two different species are strictly homologous by species descent. On the other hand, a P6-like protein and an ADH from two different species are homologous both by species descent and by locus descent from an ancestral duplication event. The available set of proteins belonging to the P6–ADH family therefore involves several sister groups of proteins. We know that the rate of amino acid replacement is different between the sister groups, because our data show that the methionine content has changed dramati-

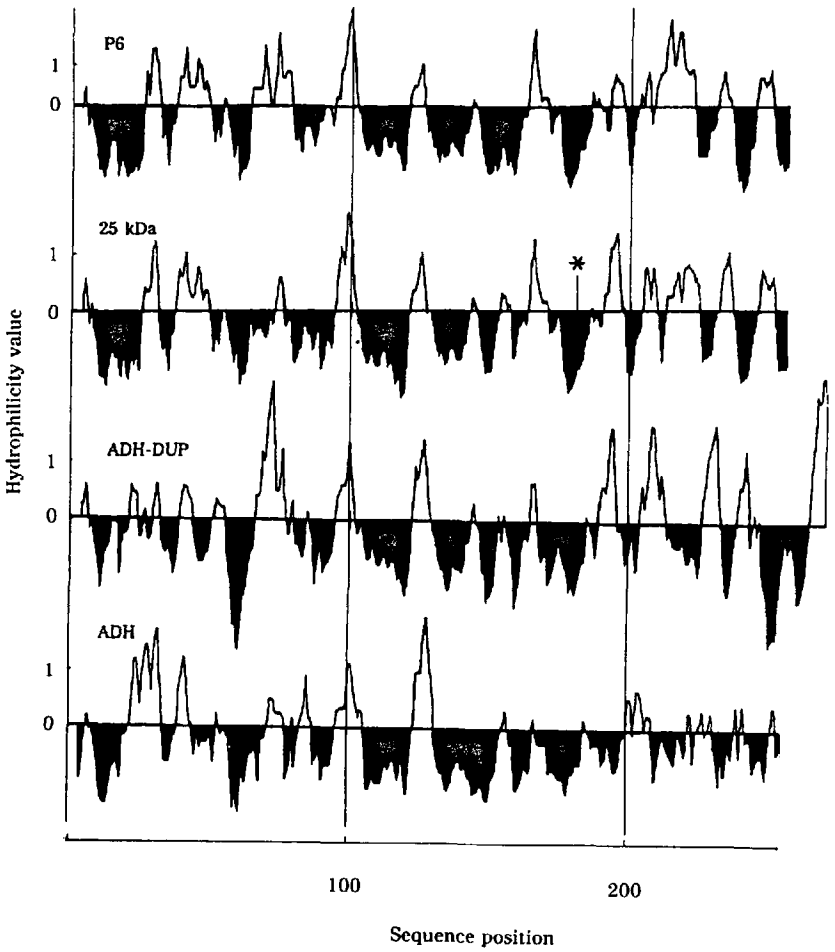

Fig. 6. Distribution of hydropathy along the amino acid sequence of P6 and three homologous insect proteins. Hydropathy was calculated using a moving average of six amino acids, according to Hopp and Woods (1981). Empty sectors, hydrophilic segments; black sectors, hydrophobic segments; star, position considered to be a stop codon by Matsumoto et al. (1985).

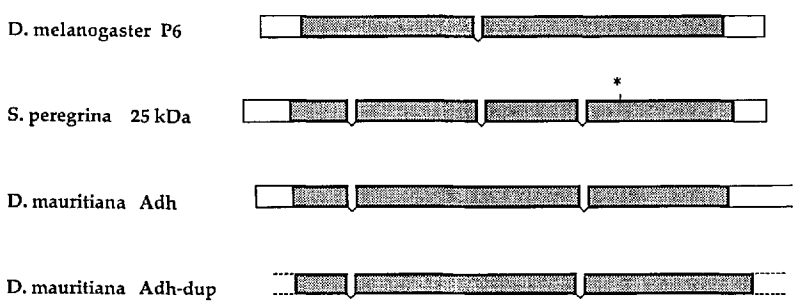

Fig. 7. Position of the introns within the ORFs of P6 and three homologous insect proteins; star, position considered to be a stop codon by Matsumoto et al. (1985)

cally. We therefore present two trees, one based on the complete number of amino acid replacements, the other on data corrected for the high turnover of methionines. The first tree is unlikely to follow a molecular clock scheme, and is only used to show the absolute amount of divergence. Only the second tree can be used to calculate a substitution rate. The scale of amino acid replacement is known in theory to always be nonlinear, as any long-term divergence must be corrected for multiple hits. Dayhoff (1978) computed the relationship between the actual number of differences observed (p_d) and the average number of substitutions per amino acid (K_{aa}) following the hypothesis of a constant change. According to Kimura (1983), this relation follows the formula $K_{aa} = -\log_e(1 - p_d - \frac{1}{5} p_d^2)$. Dayhoff's estimate being based on an average polypeptide derived from a sample where methionines accounted for only 1.1% of all amino acids, this relation is a sufficiently close approximation to be valid for the tree calculated for nonmethionine residues.

The sample of proteins we used comprised P6, the ADH-DUPs from *D. mauritiana* and *D. pseudoobscura,* the ADHs from those two species and from the more divergent species *Drosophila mulleri* (Fischer and Maniatis 1985), and 25-kd from *S. peregrina.*

Figure 8 shows both the tree obtained from absolute differences (dotted line) and the corrected tree excluding any methionine position in pairwise poly-

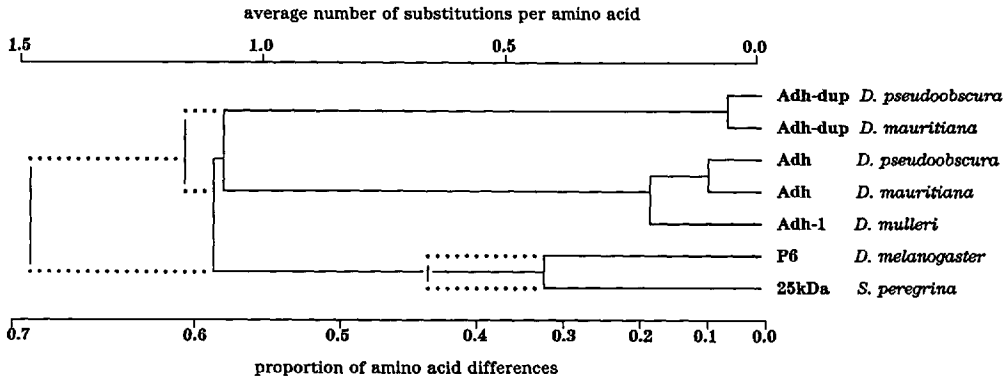

Fig. 8. UPGMA tree of amino replacements (lower scale) in the P6–ADH family of proteins. The average rate of substitution per amino acid (upper scale) was calculated according to Kimura (1983). Dotted line, tree obtained from replacements at any position; continuous line, tree corrected for methionine mutations. In the latter case, a site is excluded from the calculation of the distance between two proteins when a methionine is present in any of them.

```
                10        15        20        85        90
Liver   ADH     CAVFGLGGVGLSV
                    |   |   |
D. mau. ADH     IFVAGLGGIGLDT     LKTVDVLINGAGI
D. mau. DUP     CYVADCGGIALET     MDYIDVLINGATL
D. mel. P6      VYVGSFSGIGWQM     MGHIDVMINGEGV
S. per. 25K     VYVGGFSGFGYQV     VGHVDVLINGVGG
                    |   |   |
B. m.   GlDH    VITGSSTGLGKSM     FGKLDVMINNAGL
K. a.   RDH     AITGAASGIGLEC     TGRLDIFHANAGA
E. c.   GUT     VVIGGGQTLGAFL     FGRVDLLVYSAGI
P. p.   bphB    LITGGASGLGRAL     FGKIDTLIPNAGI
Human   17B     LITGCSSGIGLHL     EGRVDVLVCNAGL
                    |   |   |
M. m.   adip    LVTGAGKGIGRDT     IGPVDLLVNNAAL
R. m.   nodG    LVTGASGAIGGAI     LEGVDILVNNAGI
E. sp.  P-27    IITGGTRGIGFAA     TGRLDVMINNAGI
S. c.   act3    LVTGATSGIGLEI     YGPVDVLVNNAGR
```

Fig. 9. Partial alignment of four insect proteins (D.mel ADH, *D. melanogaster* ADH; D.mau DUP, *D. mauritiana* ADH-DUP; D.mel P6, *D. melanogaster* P6, S.per 25K, *Sarcophaga peregrina* 25 kd), of five short-chain dehydrogenases (B.m GlDH, *Bacillus megaterium* glucose dehydrogenase; K.a RDH, *Klebsiella aerogenes* ribitol dehydrogenase; E.c GUT, *Escherichia coli* glucitol-6-phosphate dehydrogenase; P.p bphB, *Pseudomonas pseudoalcaligenes* dihydrodiol dehydrogenase; human 17B, human 17-β-hydroxysteroid dehydrogenase), and of four related proteins (M.m adip, protein from glucocorticoid controlled gene from mouse adipocyte; R.m nodG, *Rhyzobium meliloti* nodulation protein G; E.sp P-27, *Eubacterium* sp. 27-kd protein; S.c act3, *Streptomyces coelicolor* Act III gene protein). The first line (liver ADH) displays amino acid alignment of horse liver ADH with short-chain dehydrogenases in an area corresponding to the beginning of the adenine-binding site. The second alignment corresponds to the beginning of the nicotinamide-binding site. Numbering is according to the *Drosophila* ADH mature protein. Bold letters, conserved positions in short-chain dehydrogenases. Vertical lines point to the glycine residue patterns 14–16–19 and 13–17–19, which are discussed in the text.

peptide comparisons (continuous line). Allowing for this correction, the divergence between P6 and 25-kd is reduced by one-third, and that between the P6–25-kd branch and the ADH–ADH-DUP branch is virtually abolished.

These values compare favorably with the present views on the evolution of Diptera. Paleontological and biogeographical data (Beverley and Wilson 1984; Grimaldi 1987; Lachaise et al. 1988) along with biochemical distances (Ashburner et al. 1984; Beverley and Wilson 1984) have been used by several authors to scale *Drosophila* history. The divergence of *D. mauritiana* and *D. melanogaster* from *D. pseudoobscura* is between 20 and 46 million years (Myr). Their divergence from *D. mulleri* is between 37 and 62 Myr. Their divergence from *Sarcophaga* is between 70 and 135 Myr. When taken as an order of magnitude of the boundary values of species divergence, these estimates are broadly consistent with replacement distances in the P6–ADH family of proteins. The corrected average amino acid substitution rate between P6 and 25-kd (0.43) is about twice that between the ADHs from *D. mulleri* and other species (0.22), which is itself about twice that between *D. melanogaster* and *D. pseudoobscura* ADHs (0.10). From a consideration of locus divergence on the time scale of species divergence, 25-kd and P6 are likely to be orthologous: the two genes would represent the same gene in two different species. This tree further shows that the P6–25-kd lineage was established shortly before the separation of ADH and ADH-DUP. This event cannot be dated precisely as estimations for such sequence divergence are inaccurate.

Relation to Other Short-Chain Dehydrogenases

P6 and the related insect proteins have retained substantial sequence similarity (Fig. 5): 38 amino acid residues are identical in all four proteins. These include the conservation of several arrays of amino acids such as the sequence DVxING on the D β-strand, and the sequence D(K/R)xxxGxGG upstream of the F β-strand. Several polar residues are conserved, such as acidic residues at positions 41, 86, 101, and 123, and basic residues at positions 29 and 124.

Because this could strengthen the hypothesis of

a dehydrogenase activity for P6, these four proteins were compared to nine proteins classified among the short-chain dehydrogenases (Jörnvall et al. 1981). Comparison over the entire sequence (not shown) indicated that aspartic acids 86 and 123, and tyrosine 151 are conserved on all sequences. Tyrosine 151 of P6 is homologous to tyrosine 160 of ribitol dehydrogenase, which is involved in subunit binding (Jany et al. 1984). Two areas of substantial similarity are displayed in Fig. 9. One is downstream from the A β-strand, and the other is homologous to the sequence DVxING reported above from the D β-strand. The first of these two sequences is usually considered critical in the enzyme activity of NAD^+ enzymes (Wierenga et al. 1985). The fingerprint sequence xGxGxxG has been proposed (Wierenga et al. 1985) for this region. Vertebrate ADHs are similar to *D. melanogaster* Adh in this respect (Benyajati et al. 1981; Scrutton et al. 1990), with the sequence xGLGGxG, as read from position 13 of the mature enzyme of *D. melanogaster* (Fig. 9). The critical function of the glycine at position 14 of *Drosophila* ADH is supported by the fact that its replacement by aspartic acid is considered to be responsible for the inactivation of the ADH-N^{11} mutation of *D. melanogaster* (Benyajati et al. 1981). Glycine is rarely replaced in coenzyme-binding domains, probably because this small-sized residue allows the coenzyme to be accommodated within the protein folding (Jörnvall et al. 1984). Experimental substitution of glycine 14 of *Drosophila* ADH by either alanine or valine results in a decrease in its activity (Chen et al. 1990). However, the ADH of *Drosophila lebanonensis* (Villarroya et al. 1989) does have an alanine at this position. In contrast to the view that glycines 14, 16, and 19 of ADH play an important role, P6 and most of the other short-chain dehydrogenases (Fig. 9) have a glycine at the corresponding positions 13, 17, and 19, resulting in the fingerprint GxxxGxG. This pattern is probably very ancient, as it remained unchanged in bacteria, mammals, and insects. However, it changed drastically along the evolutionary lineage leading to ADH from its common ancestor with P6.

The divergence of P6 and Adh has clearly resulted from a duplication event that took place long enough ago for the two genes to have substantially diverged from their common ancestor. This is an example of the role played by gene duplication in providing new material for further functional diversification (Ohno 1970). This makes P6 a valuable tool for examining the evolutionary convergence that resulted in a short-chain dehydrogenase evolving a metabolic function analogous to plant and mammal ADHs.

Acknowledgments. We thank Michael Ashburner and François Brégégère for their advice and discussion. We are grateful to Martin Kreitman for communicating his unpublished results and to Helen Benes, Matthew Cobb, and Andrea Nicholson for helpful criticism and careful reading of the manuscript. This work was supported by grants from the Centre National de la Recherche Scientifique and the Institut National de la Santé et de la Recherche Médicale (grant no. 841017) to J.-A.L.

References

Ashburner M, Bodmer M, Lemeunier F (1984) On the evolutionary relationships of *Drosophila melanogaster*. Dev Genet 4:295–312

Baker ME (1990) A common ancestor for human placental 17β-hydroxysteroid dehydrogenase, *Streptomyces coelicolor* actIII protein, and *Drosophila melanogaster* alcohol dehydrogenase. FASEB J 4:222–226

Benyajati C, Place AR, Powers DA, Sofer W (1981) Alcohol dehydrogenase gene of *Drosophila melanogaster*: relationship of intervening sequences to functional domains in the protein. Proc Natl Acad Sci USA 78:2717–2721

Beverley SM, Wilson AC (1984) Molecular evolution in *Drosophila* and the higher Diptera II. A time scale for fly evolution. J Mol Evol 21:1–13

Birnstiel ML, Busslinger M, Strub K (1985) Transcription termination and 3′ processing: the end is in site! Cell 41:349–359

Calzone FJ, Britten RJ, Davidson EH (1987) Mapping of gene transcripts by nuclease protection assays and cDNA primer extension. In: Berger SL, Kimmel AR (eds) Guide to molecular cloning techniques. Academic, London, pp 611–632

Chambers GK (1988) The *Drosophila* alcohol dehydrogenase gene–enzyme system. Adv Gene 25:39–107

Chen Z, Lu L, Shirley M, Lee WR, Chang SH (1990) Site-directed mutagenesis of glycine-14 and two "critical" cysteinyl residues in *Drosophila* alcohol dehydrogenase. Biochemistry 29:1112–1118

Cherbas L, Schulz RA, Koehler MD, Savakis C, Cherbas P (1986) Structure of the *Eip28/29* gene, an ecdysone-inducible gene from *Drosophila*. J Mol Biol 189:617–631

Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol 47:45–148

Cohn VH, Moore GP (1988) Organization and evolution of the alcohol dehydrogenase gene in *Drosophila*. Mol Biol Evol 5: 154–166

Coleman JP, White WB, Lijewski M, Hylemon PB (1988) Nucleotide sequence and regulation of a gene involved in bile acid 7-dehydroxylation by *Eubacterium* sp. strain VPI 12708. J Bacteriol 170:2070–2077

David JR (1977) Signification d'un polymorphisme enzymatique: la déshydrogénase alcoolique chez *Drosophila melanogaster*. Ann Biol 16:451–472

Dayhoff MO (1978) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington DC

Debellé F, Sharma SB (1986) Nucleotide sequence of *Rhizobium meliloti* RCR2011 genes involved in host specificity of nodulation. Nucleic Acids Res 14:7453–7472

Delaney SJ, Smith DF, McClelland A, Sunkel C, Glover DM (1986) Sequence conservation around the 5′ ends of the larval serum protein 1 genes of *Drosophila melanogaster*. J Mol Biol 189:1–11

Deutsch J, Laval M, Lepesant JA, Maschat F, Pourrain F, Rat L (1989) Larval fat body-specific gene expression in *D. melanogaster*. Dev Genet 10:220–231

Fischer JA, Maniatis T (1985) Structure and transcription of

the *Drosophila mulleri* alcohol dehydrogenase genes. Nucleic Acids Res 13:6899–6917

Furukawa K, Arimura N, Miyazaki T (1987) Nucleotide sequence of the 2,3-dihydroxybiphenyl dioxygenase gene of *Pseudomonas pseudoalcaligenes*. J Bacteriol 169:427–429

Grantham R, Perrin P, Mouchiroud D (1986) Patterns in codon usage of different kinds of species. In: Dawkins R, Ridley M (eds) Oxford surveys in evolutionary biology. Oxford University Press, Oxford, pp 48–81

Grimaldi DA (1987) Amber fossil Drosphilidae (Diptera), with particular reference to the hispaniolan taxa. Am Mus Novit 2880:1–23

Hallam SE, Malpartida F, Hopwood DA (1988) Nucleotide sequence, transcription and deduced function of a gene involved in polyketide antibiotic synthesis in *Streptomyces coelicolor*. Gene 74:305–320

Heberlein U, England B, Tjian R (1985) Characterization of *Drosophila* transcription factors that activate the tandem promoters of the alcohol dehydrogenase gene. Cell 41:965–977

Henikoff S (1984) Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene 23:351–359

Henikoff S, Wallace JC (1988) Detection of protein similarities using nucleotide sequence databases. Nucleic Acids Res 16: 6191–6204

Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78:3824–3828

Jany KD, Ulmer W, Fröschle M, Pfleiderer G (1984) Complete amino acid sequence of glucose dehydrogenase from *Bacillus megaterium*. FEBS Lett 165:6–10

Jörnvall H, Persson M, Jeffery J (1981) Alcohol and polyol dehydrogenases are both divided into two protein types, and structural properties cross-relate the different enzyme activities within each type. Proc Natl Acad Sci USA 78:4226–4230

Jörnvall H, von Bahr-Lindström H, Jany KD, Ulmer W, Fröschle M (1984) Extended superfamily of short alcohol-polyol-sugar dehydrogenases: structural similarities between glucose and ribitol dehydrogenases. FEBS Lett 165:190–196

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, pp 55–97

Kozak M (1986) Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. Proc Natl Acad Sci USA 83:2850–2854

Kreitman M (1987) Molecular population genetics. In: Harvey PH, Partridge L (eds) Oxford surveys in evolutionary biology. Oxford University Press, Oxford, pp 38–60

Krook M, Marekov L, Jörnvall H (1990) Purification and structural characterization of placental NAD^+-linked 15-hydroxyprostaglandin dehydrogenase. The primary structure reveals the enzyme to belong to the short-chain alcohol dehydrogenase family. Biochemistry 29:738–743

Lachaise D, Cariou ML, David JR, Lemeunier F, Tsacas L, Ashburner M (1988) Historical biogeography of the *Drosophila melanogaster* species subgroup. In: Hecht MK, Wallace B, Prance GT (eds) Evolutionary Biology, vol 22. Plenum, New York, pp 159–225

Lepesant JA, Levine M, Garen A, Kejzlarvoa-Lepesant J, Rat L, Somme-Martin G (1982) Developmentally regulated gene expression in *Drosophila* larval fat bodies. J Mol Appl Genet 1:371–383

Maniatis T, Hardison RC, Lacy E, Lauer J, O'Connell C, Quon D, Sim GK, Efstratiadis A (1978) The isolation of structural genes from libraries of eucaryotic DNA. Cell 15:687–701

Marekov L, Krook M, Jörnvall H (1990) Prokaryotic 20β-hydroxysteroid dehydrogenase is an enzyme of the "short-chain, non-metalloenzyme" alcohol dehydrogenase type. FEBS Lett 266:51–54

Maschat F, Dubertret ML, Therond P, Claverie JM, Lepesant JA (1990) Structure of the ecdysone-inducible *P1* gene of *Drosophila melanogaster*. J Mol Biol 214:359–372

Matsumoto N, Sekimizu K, Soma GI, Ohmura Y, Andoh T, Nakanishi Y, Obinata M, Natori S (1985) Structural analysis of a developmentally regulated 25-kDa protein gene of *Sarcophaga peregrina*. J Biochem 97:1501–1508

Navre M, Ringold G (1988) A growth factor-repressible gene associated with protein kinase C-mediated inhibition of adipocyte differentiation. J Cell Biol 107:279–286

Ohno S (1970) Evolution by gene duplication. Springer Verlag, Berlin

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, ed 2. Cold Spring Harbor Laboratories, Cold Spring Harbor NY

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467

Schaeffer SW, Aquadro CF (1987) Nucleotide sequence of the *Adh* gene region of *Drosophila pseudoobscura*: evolutionary change and evidence for an ancient gene duplication. Genetics 117:61–73

Scrutton NS, Berry A, Perham RN (1990) Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. Nature 343:38–43

Shen NLL, Subrahmanyam G, Clark W, Martin P, Sofer W (1989) Analysis of *Adh* gene regulation in *Drosophila*: studies using somatic transformation. Dev Genet 10:210–219

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Simon I, Cserzö M (1990) Rapid evolution of the amino acid composition of proteins. Trends Biochem Sci 15:135–136

Sofer W, Martin PF (1987) Analysis of alcohol dehydrogenase gene expression in *Drosophila*. Annu Rev Genet 21:203–225

Sullivan DT, Atkinson PW, Starmer WT (1990) Molecular evolution of the alcohol dehydrogenase genes in the genus *Drosophila*. Evol Biol 24:107–147

Villarroya A, Juan E, Egestad B, Jörnvall H (1989) The primary structure of alcohol dehydrogenase from *Drosophila lebanonensis*. Eur J Biochem 180:191–197

Wierenga RK, De Maeyer MCH, Hol WGJ (1985) Interaction of pyrophospate moieties with α-helixes in dinucleotide binding proteins. Biochemistry 24:1346–1357

Yamada M, Saier MH Jr (1987) Glucitol-specific enzymes of the phosphotransferase system in *Escherichia coli*. J Biol Chem 262:5455–5463

Yokoyama S, Yokoyama R, Kinlaw CS, Harry DE (1990) Molecular evolution of the zinc-containing long-chain alcohol dehydrogenase genes. Mol Biol Evol 7:143–154

Received December 26, 1990/Revised April 1, 1991