Molecular Structure and Transformation of the Glucose Dehydrogenase Gene in *Drosophila melanogaster*

Ross Whetten,[1] Edward Organ, Philip Krasney, Diana Cox-Foster[2] and Douglas Cavener[3]

Department of Molecular Biology, Vanderbilt University, Nashville, Tennessee 37235

Manuscript received February 15, 1988
Revised copy accepted June 10, 1988

ABSTRACT

We have precisely mapped and sequenced the three 5′ exons of the *Drosophila melanogaster Gld* gene and have identified the start sites for transcription and translation. The first exon is composed of 335 nucleotides and does not contain any putative translation start codons. The second exon is separated from the first exon by 8 kb and contains the *Gld* translation start codon. The inferred amino acid sequence of the amino terminus contains two unusual features: three tandem repeats of serine-alanine, and a relatively high density of cysteine residues. *P* element-mediated transformation experiments demonstrated that a 17.5-kb genomic fragment contains the functional and regulatory components of the *Gld* gene.

THE glucose dehydrogenase gene (*Gld*) in *Drosophila melanogaster* is required at a single stage in development for the modification of the puparium. *Gld* mutants fail to eclose at the termination of metamorphosis but can be easily rescued by excising the anterior end of the puparium case (CAVENER and MACINTYRE 1983). Despite the simple mutant phenotype, the GLD enzyme and mRNA are transiently expressed at every major stage of development (CAVENER *et al.* 1986a; CAVENER 1987a). The temporal pattern of *Gld* mRNA accumulation is highly correlated with accumulation of the major molting hormone ecydsterone and has been demonstrated to be regulated by this hormone during the third larval instar (M. MURTHA and D. CAVENER, unpublished data). *Gld* mRNA is expressed in a variety of ectodermal tissues including the hypophysis and antennal-maxillary complex (embryos); the anterior spiracular gland cells and the epidermis (third instar larvae); wings, legs, antennae, cibarium, epidermis, rectal papillae, neck, trachea, and some components of the reproductive tract of both male and female (pharate adults) (D. FOSTER-COX, C. SCHONBAUM and D. CAVENER, unpublished results). During the adult stage *Gld* expression is almost entirely limited to the male anterior ejaculatory duct (CAVENER and MACINTYRE 1983; CAVENER *et al.* 1986a). The GLD enzyme is apparently secreted since it can be recovered from the molting fluid of pupae and is transferred from adult males to females during copulation.

Inasmuch as *Gld* regulation involves steroid hormone control, a germ layer lineage restriction, and sexual differentiation, it is an excellent paradigm for developmentally regulated genes. In order to elucidate the *cis*-acting elements which are responsible for the various aspects of its regulation, we have engaged in a detailed molecular analysis of the *Gld* gene. Genomic DNA clones of the *Gld* gene were isolated by the method of chromosome walking, and identified on the basis of the localization of three independent *Gld* mutations (CAVENER *et al.* 1986a). These three mutations were found in a 7-kb region at 84C8 on the right arm of the third chromosome. Northern hybridization analysis identified a 2.8-kb poly(A^+) RNA derived from this region which is highly correlated with the pattern of expression exhibited by the GLD enzyme throughout development. This correlation includes the virtual restriction during the adult stage to the male ejaculatory duct. We describe fine-scale mapping of the 5′ half of the gene including the start site of transcription and translation. Proof that a 17.5-kb restriction fragment including the transcription unit contains the entire *Gld* gene is provided by *P* element-mediated gene transformation experiments.

MATERIALS AND METHODS

Subcloning and DNA sequencing: For the analysis of exon I genomic restriction fragments from lambda clone E14b (CAVENER *et al.* 1986a) were inserted into Bluescript KS(+) or KS(−) phagemids (Stratagene, Inc.) which contain T7 and T3 phage promoters and can be propagated as a double-stranded DNA plasmid or, upon superinfection with helper phage, as a single-stranded DNA phage. The genomic subclones used for the analysis of exons II and III were previously described (CAVENER *et al.* 1986a). These subclones are in the SP64/65 vectors (Promega Biotec, Inc.), which contain the SP6 phage promoter, or in pEMBL8/9 phagemid vectors (DENTE, CESARENI and CORTESE 1983).

[1] Present address: Department of Biology, Utah State University, Logan, Utah 84322.
[2] Present address: Department of Entomology, Pennsylvania State University, University Park, Pennsylvania 16802.
[3] To whom correspondence should be addressed.

A series of terminal deletions were constructed from the subclones in the phagemid vectors using the HENIKOFF (1984, 1987) *Exo*III method. These deletions were used for DNA sequencing and transcript mapping experiments. DNA sequencing of single strand templates from the deletion mutants was performed by the chain termination method (SANGER and COULSON 1975) using the Klenow fragment of *Escherichia coli Pol*I and [α-^{35}S]dATP. In some cases double stranded templates were sequenced using the alkaline denaturation method of CHEN and SEEBURG (1985). The DNA sequences of exons I, II and III were verified by sequencing each nucleotide from two independent DNA templates or (in most cases) by sequencing both strands.

Northern hybridization: Total RNA and poly(A$^+$) RNA were isolated after previously published procedures (CAVENER *et al.* 1986a). The RNAs were fractionated on 2.2 M formaldehyde/1.2% agarose gels (SEED 1982). RNA gels were blotted to nitrocellulose or nylon membranes. The cRNA probes were prepared following the procedures of Promega Biotec. Hybridizations were performed at 58°C in standard hybridization buffer containing 50% (v/v) formamide for 15–24 hr. Filters were washed at room temperature with 2 × SSC–0.2% (w/v) SDS and then at 65° with 0.2 × SSC–0.2% (w/v) SDS.

RNAse protection experiments: The nuclease protection procedure of ZINN, DIMAIO and MANIATIS (1983) was initially used to map exons II and III. Poly(A$^+$) RNA (1–5 μg) precipitated with ethanol was redissolved in 28 μl of hybridization buffer (80% (v/v) formamide, 400 mM NaCl, 40 mM PIPES, pH 6.4). RNA probes were synthesized using 50–75 μCi of [α-^{32}P]UTP and 1 μg of DNA template. After synthesis of the cRNA probe, the DNA template was removed by digestion with DNase I, the reaction extracted with phenol and chloroform, and the probes recovered by ethanol precipation. The probes were redissolved in 20–50 μl of hybridization buffer, and 2 μl were then added to the RNA solution. The hybridization mixture was heated to 80° for 3 min, then hybridized overnight at 40–50°. The hybrids were then digested in 300 μl of an RNAse solution (40 μg/ml RNAse A, 2 μg/ml RNAse T1, 300 mM NaCl, 10 mM Tris pH 7.5, 5 mM EDTA) for 1 hr at 23°. The digestions were stopped by phenol/chloroform extraction and the hybrids precipitated by the addition of 5 μg of carrier RNA and 2 volumes of cold ethanol. The samples were electrophoresed on sequencing gels for maximum resolution. FISCHER and MANIATIS (1985) had noted that high concentrations of RNase A can lead to undesired cleavage of poly(U)-poly(A) hybrid tracts. Since exon I contains such a sequence we found that it was necessary to reduce the concentration of RNase A in the reaction by 400-fold (*i.e.*, to 0.1 μg/ml).

Primer extension experiments: The method used for primer extension experiments was modified from a protocol obtained from ROBERT THOMPSON (personal communication). Oligonucleotides were synthesized with a Biosearch DNA synthesizer in the laboratory of STEVEN LLOYD (Vanderbilt University). The oligonucleotide primers were 5′-end-labeled using [γ-^{32}P]ATP (MANIATIS, FRITSCH and SAMBROOK 1982). Poly(A$^+$) RNA (1–5 μg) was mixed with 5 pmol of 5′-end-labeled oligonucleotide primer in a total volume of 10 μl of distilled, pyrocarbonic acid diethyl ester-treated water. The mixture was heated 3 min at 65° and then incubated for 30–60 min at 43° or 50°. An equal volume of 2× reaction buffer was then added to start the extension reaction. The 2× reaction buffer contained 100 mM Tris (pH 8.3), 20 mM DTT, 20 mM $MgCl_2$, 2 mM of each of the four dNTPs, 100 units/ml placental RNAse inhibitor and 1000 units/ml avian myeloblastosis virus reverse transcriptase. The extension reaction was allowed to proceed 30 min at 43°, then stopped by the addition of EDTA to a final concentration of 20 mM. Sodium acetate was added to a final concentration of 0.3 M and the reaction products precipitated with ethanol. The redissolved samples were electrophoresed on sequencing gels.

***P* element-mediated transformation:** The pWG67 transformation clone was constructed from a 9.7-kb *Kpn*I-*Sal*I fragment containing the 5′ half of *Gld*, a 7.8-kb *Sal*I-*Kpn*I fragment containing the 3′ half of *Gld*, and the pW5 *P* element-transformation vector of KLEMENZ, WEBER and GEHRING (1987). The pWG67 *Gld* clone was coinjected along with the pπ25.7wc helper *P* element into *y w* (yellow; white) preblastoderm embryos using the general procedures of SPRADLING and RUBIN (1982). Survivors were backcrossed to the *y w* host strain and putative transformants among the progeny were detected as adult flies with red eyes and yellow bodies. As noted by KLEMENZ *et al.* (1987) transformants using the pW5 vector do not typically exhibit wild-type red eye color; instead they display colors similar to *w* hypomorphs. Two independently transformed strains, T-7.1 and T-14.3, were found to carry pWG67 inserts on chromosome *II*. T-7.1 and T-14.3 were separately crossed to a *Gld* null strain (*w/w; Gld*^{n1}/*Gld*^{n2}*cu/cu*). F_1 males were backcrossed to the *Gld* null parent strain and their progeny analyzed. As expected red eye-curled wing (*w*$^+$ *cu*) flies self-eclosed whereas white eye-curled wing (*w cu*) flies died in the "head-jammed" state typical of the *Gld* lethal phenotype. Approximately equal numbers of these two phenotypes were observed, consistent with independent assortment of the rescuing factors (*i.e.*, the pWG67 insertions) on chromosome *II* from the curled mutation (tightly linked to *Gld*) on chromosome *III*. The T-14.3 insertion induced a recessive lethal mutation so it has not been possible to create a homozygous strain. Genomic Southern analysis of the transformants was performed after previously published procedures (CAVENER *et al.* 1986a) to verify the integration of the transforming DNA and to determine copy number.

RESULTS

Low resolution mapping of the *Gld* 2.8-kb mRNA: A series of 22 restriction fragments from a 27-kb region containing the *Gld* gene was subcloned into one of three vectors which support the synthesis of cRNA probes. Single stranded cRNA probes representing both strands were used to probe Northern blots containing total RNA from pharate adults and female and male adults. A summary of these results is presented in Figure 1. Probes spanning a region of 14 kb of genomic DNA detected the *Gld* 2.8-kb mRNA. The positive probes were clustered in three groups separated by two putative intronic regions. The data germane to the 3′ half of the *Gld* transcription unit have been published (CAVENER *et al.* 1986a). Only one other RNA species, 1.5 kb in length, has been consistently observed to hybridize to probes from this region. The 1.5-kb RNA is identified by a few of the 3′ probes which also identify the 2.8-kb RNA. The temporal pattern of expression of the 1.5-kb RNA is not correlated with the pattern of GLD enzyme expression. Two RNAs complementary to the strand containing *Gld* mRNA were detected using

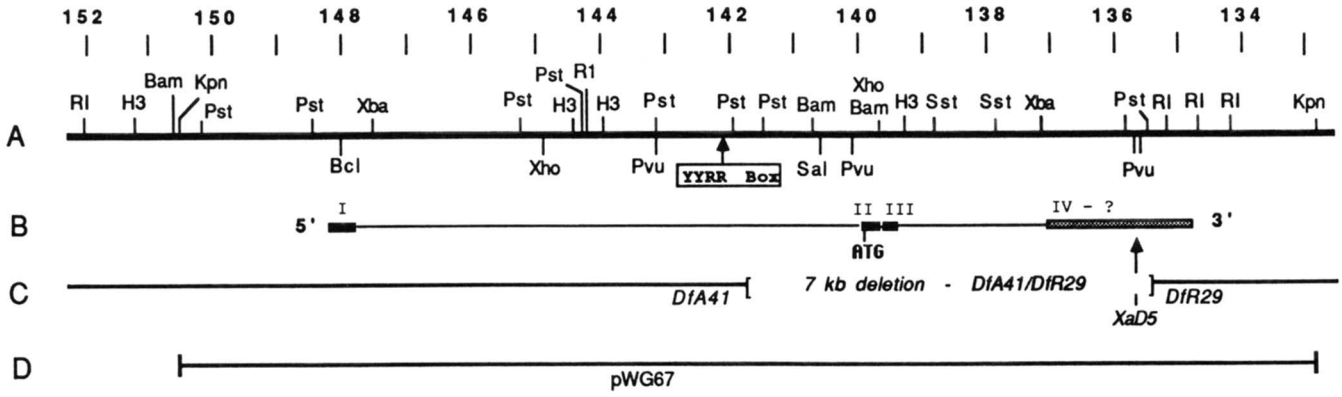

FIGURE 1.—Molecular map of the *Gld* gene. (A) Restriction map derived from genomic clones and partially confirmed by whole genome Southern analysis. (B) Transcript map of the 2.8-kb *Gld* mRNA. Thick lines and numbers indicate the location and order of exons. Stippled line denotes the region of the 3′ exons (IV–?) which have not been precisely mapped. Thin lines denote introns. (C) Composite 7-kb deletion *DfA41/DfR29* and the T2;3 *XaD5* reciprocal translocation breakpoint. As predicted, flies bearing these mutations lack GLD enzyme activity, lack the *Gld* 2.8-kb mRNA, and exhibit the noneclosion mutant phenotype characteristic of *Gld* mutants. (D) The 17.5-kb *Kpn*I fragment within the pWG67 clone which is able to provide *Gld* functions upon transformation.

probes from this region. However, these antisense RNAs are not transcribed from the *Gld* locus (CAVENER *et al.* 1986a).

The 5′ most genomic probes which hybridize to the *Gld* 2.8-kb RNA map more than 5 kb upstream from the composite A41/R29 deletion which genetically localizes *Gld*. In order to provide further evidence that the 2.8-kb RNA detected by the 5′ extreme probes corresponds to the *Gld* mRNA, a Northern blot containing RNA from the A41/R29 deletion was probed with a complementary radiolabeled cRNA from this region (Figure 2). As predicted the 2.8-kb RNA is not detected in the A41/R29 deletion, as was previously shown for probes corresponding to more 3′ exonic regions (CAVENER *et al.* 1986a). The Northern blot was subsequently hybridized with an Adh (alcohol dehydrogenase) probe to confirm the presence and integrity of the RNA in the A41/R29 sample.

High resolution mapping: Exon I—The genomic subclone pCG4, containing a 4.2-kb *Xba*I/*Eco*RI fragment, was determined to correspond to the putative 5′ end of the *Gld* gene (coordinates 147.4–151.6 of Figure 1) Two sets of terminal deletion mutations of pCG4 (Figure 3) were constructed using the HENIKOFF (1984) *Exo*III method. One set contained deletions from the *Xba*I end while the other set contained deletions from the *Eco*RI end (Figure 3). Single-stranded cRNA probes were made from representatives from each of the two sets and used to probe Northern blots of poly(A$^+$) or total RNA from pharate adults. These experiments lead to the delineation of a region between the deletion breakpoints of subclones 25a and 31a which hybridize with the *Gld* 2.8-kb RNA.

RNase protection experiments were used to precisely map the position of exon I (Figure 4). Three

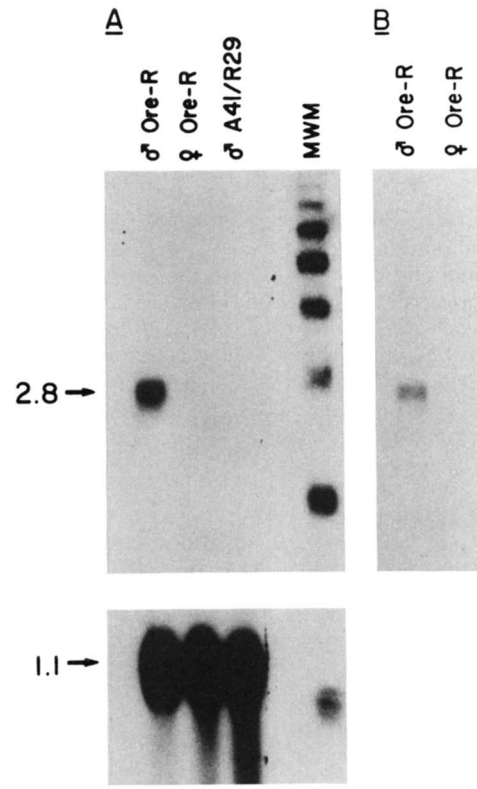

FIGURE 2.—Confirmation of the identity of the *Gld* 2.8-kb RNA. (A) A Northern blot was hybridized with a cRNA probe corresponding to the 5′ most region in Figure 1 which hybridizes with a 2.8-kb RNA. As expected the 2.8-kb RNA is not detected in A41/R29 males. This blot was reprobed with an Adh-specific radiolabeled probe to demonstrate the integrity of the A41/R29 RNA sample. The 1.1-kb RNA corresponds to the Adh mRNA. (B) A Northern blot hybridized with the ^{32}P-end-labeled 24-mer oligonucleotide used in the primer extension experiments (Figure 6).

radiolabeled cRNA probes complementary to the *Gld* mRNA were individually hybridized with pharate adult poly(A$^+$) RNA, subjected to RNase A and T1 digestion, and fractionated on urea-PAGE gels to

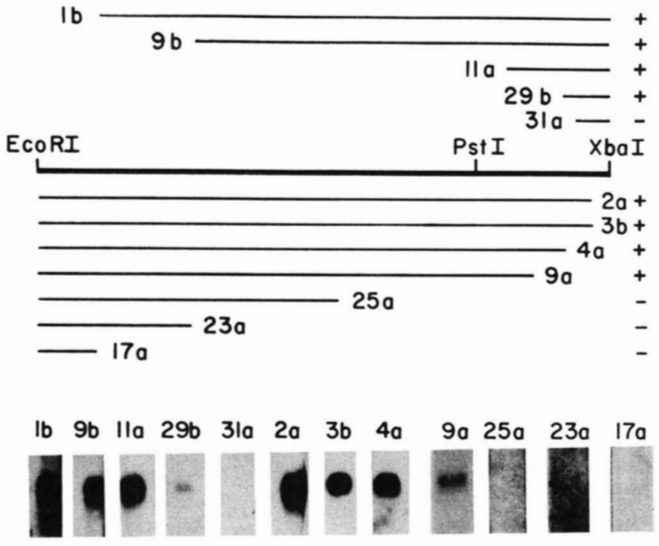

FIGURE 3.—Northern hybridization localization of exon I. (A) 5′ deletions (above the map) and 3′ deletions (below the map) of the *Eco*RI-*Xba*I fragment were constructed using the Henikoff (1984) *Exo*III method. The 4.8-kb *Eco*RI-*Xba*I fragment is shown in Figure 1 (coordinates 147.4–152.2) to contain exon I. (B) Radiolabeledd cRNA probes from each of the deletions were used to probe Northern blots containing pharate adult/adult male RNA. The autoradiograms were exposed for variable periods in order to detect weak hybridization signals. The "+" and "−" signs (A) indicate the presence and absence, respectively, of hybridization signal from the *Gld* 2.8-kb mRNA as observed in the autoradiograms (B).

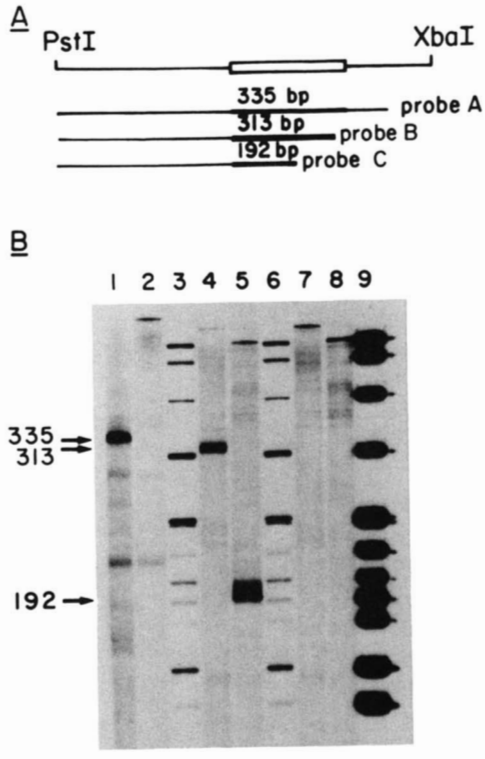

FIGURE 4.—RNAse protection mapping of exon I. (A) The final derived map of exon I is shown as an open box. The complementary cRNA probes used to determine the size and boundaries of exon I are below the map. The thick portion of each probe map represents the protected fragment after RNase digestion. (B) Lanes 1, 4 and 5 contain the fragments protected by 5 μg of pharate adult poly(A$^+$); lanes 2, 7, and 8 contain fragments protected by 5 μg of yeast RNA as a negative control; and lanes 3, 6 and 9 contain pBR322/*Hpa*II molecular weight markers. Lanes 1 and 2, probe A; lanes 4 and 7 probe B; and lanes 5 and 8 probe C.

determine the size of the protected fragments. All three experiments yielded single protected fragments. The difference in the length of the fragments protected by probes B and C is approximately equal to the difference in the lengths of the two probes suggesting that the two probes both protect the 5′ end of exon I but neither protects the 3′ end. Furthermore, this suggests a precise position of the 5′ end of exon I (*i.e.*, 313 and 192 nucleotides (nt) from the deletion breakpoints of probes B and C, respectively). Probe A is thought to protect the entire exon as implied by the deletion subclone/Northern blot analysis described above (Figure 3). Thus, the probe A-protected fragment, 335 nt, is the estimated size for exon I. Because the three probes have one common end we interpret these results as indicating the presence and position of a single 335 nt exon. It should be noted that it was necessary to modify the RNase protection protocol of ZINN, DIMAIO and MANIATIS (1983) as suggested by FISCHER and MANIATIS (1985) in order to prevent cleavage at a poly(A) tract within the *Gld* mRNA when hybridized to the complementary probes.

The DNA sequence of an 1143-bp region beginning at the Xba end of the fragment was determined (Figure 5) using the method of SANGER and COULSON (1975), as described in MATERIALS AND METHODS. A sequence (ATC/GTAAGT) similar to the 5′ splice junction consensus (MOUNT 1982) was found at the predicted position based upon the RNase protection experiments. At the predicted 5′ end of exon I is the sequence TGAGTCGG which is very similar to the Drosophila transcription start site consensus sequence (SNYDER *et al.* 1982; CHERBAS *et al.* 1986).

In order to confirm that the 5′ end of exon I corresponds to the 5′ end of the *Gld* mRNA and the putative start site of transcription, primer extension experiments were conducted. The 24-mer oligonucleotide primer used for this experimnet should bind to the *Gld* mRNA approximately 100 nt downstream of the 5′ end of exon I predicted by the RNase protection experiments. Using this primer, adult male and pharate adult poly(A$^+$) RNA yielded a major 100 ± 1 nt primer extension product (Figure 6). As predicted adult female poly(A$^+$) RNA yields very little of this extension product. These results were consistently obtained in three separate experiments. Raising the temperature of the primer extension reaction substantially reduces the amount of other primer extension products without decreasing the signal from the major 100-nt product. Therefore, we speculate that these other primer extension products are the result of random binding of the primer to other RNAs under

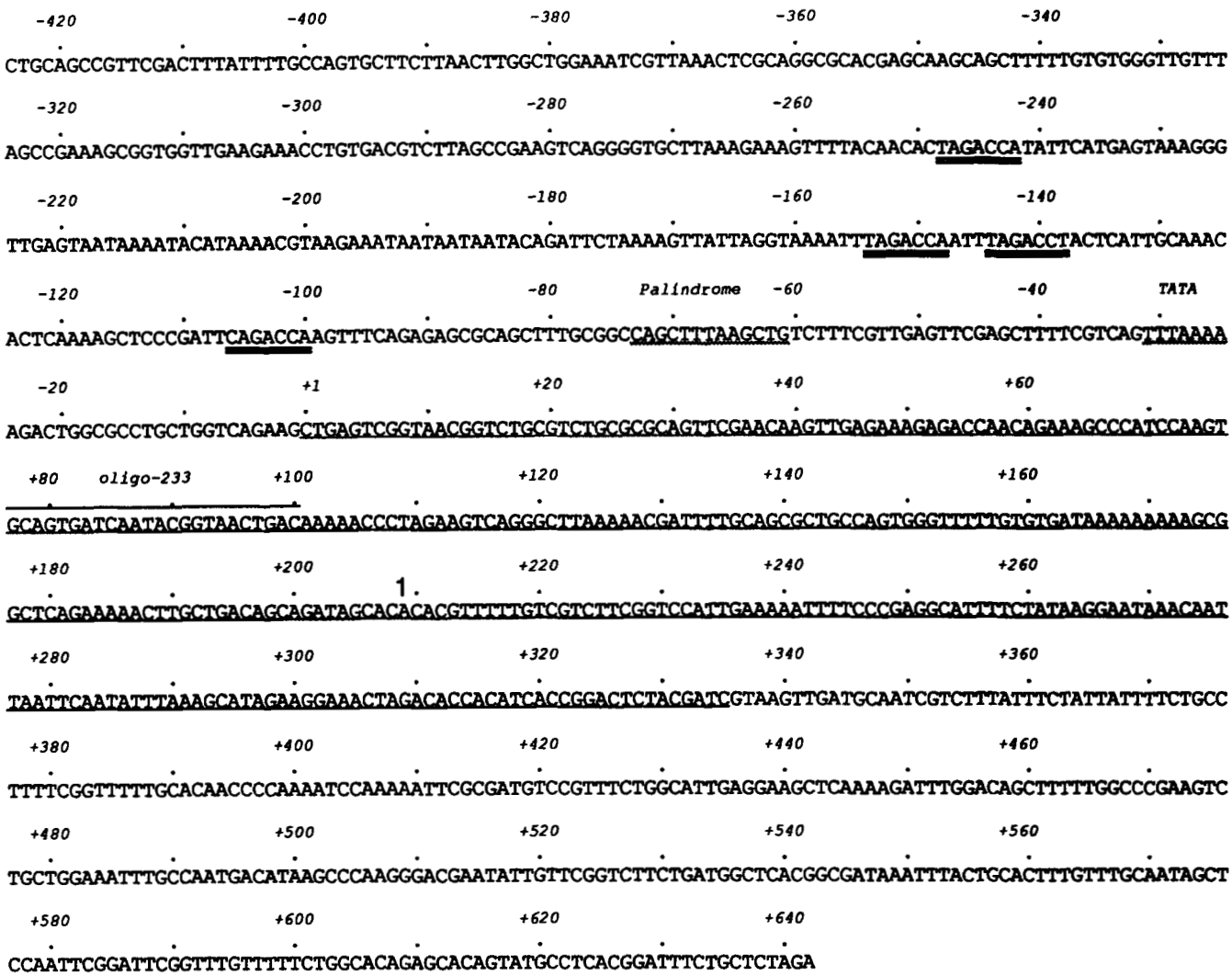

FIGURE 5.—DNA sequence of the *Gld* promoter region and exon I. The sequence begins at the *Pst*I site and ends at the *Xba*I site shown in Figure 4A and in Figure 1 (coordinates 147.5–148.6). Numbers are relative to the start site of transcription (+1). Double underlined: four direct repeats of the TAGACCA motif. Dash underline: a 13 nt palindrome (starting at −73) and the TATA box (starting at −31). Solid underline: exon I. Over line (+77 to +100): oligo-233 complementary sequence used for primer extension experiments.

the rather low stringency conditions dictated by the primer extension reaction. It is important to note that this primer detects only the *Gld* 2.8 mRNA in Northern hybridization experiments which are conducted under much higher stringency conditions (Figure 2B).

Exons II and III—Eight kilobases downstream of exon I, probes from two small adjacent restriction fragments (*Pvu*II-*Bam*HI, 450 bp; *Bam*HI-*Hin*dIII, 340 bp; see Figure 1) detect the *Gld* 2.8 mRNA (CAVENER *et al.* 1986a). The DNA sequence of this region was determined and analyzed for potential RNA splice sites and coding regions (Figure 7). From this analysis emerged a model for two small coding exons (177 nt and 121 nt) separated by a 73-nt intron (Figure 8). The details of this model were tested by RNase protection experiments. The probes used and the predicted products are shown in Figure 8A and the results of the experiments to confirm the model are shown in Figure 8, B–D. Probe A protects two RNA fragments with the predicted lengths of the two exons (Figure 8B). A cRNA probe derived from a partially digested template at the *Bam*HI site yields the 121-nt product, a small amount of the 177-nt, and a 27-nt product (not shown) which corresponds to the 3′ terminus of exon II (Figure 8C). Since probe B is predicted to terminate in exon II it should yield a single product which should confirm the position of exon II in the DNA sequence. As shown in Figure 8D the probe B experiment gives a 155-nt protected fragment. Although similar experiments were not done to determine the precise position of exon III, we are confident that our assignment of its position in Figure 7 is correct since the predicted 40 amino acids encoded in exon III are perfectly conserved within three divergent Drosophila species whereas the putative intronic sequences immediately flanking exon III

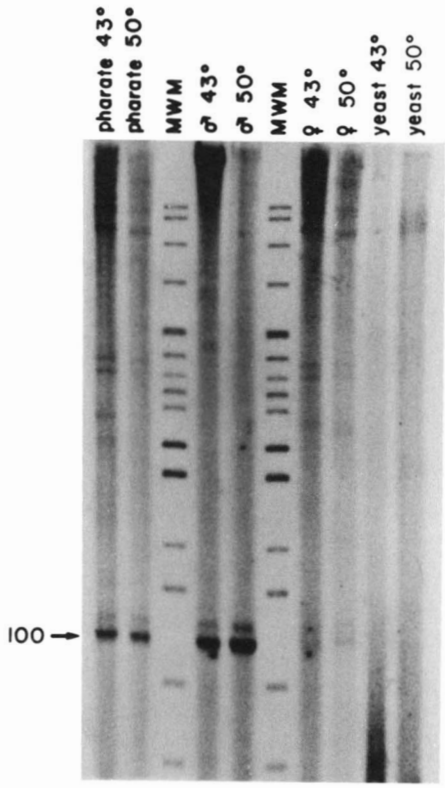

FIGURE 6.—Primer extension mapping of the 5′ end of the *Gld* mRNA. Aliquots of 5 μg of poly(A⁺) RNA from pharate adults, adult males or adult females were annealed at either 43° or 50° with oligo-233 (a 24-mer oligonucleotide). Arrow points to the major 100 nt product. See Figure 5 for sequence and position of oligo-233. Radiolabeled primer extension products were subjected to denaturing PAGE. Total yeast RNA served as the negative control for these experiments.

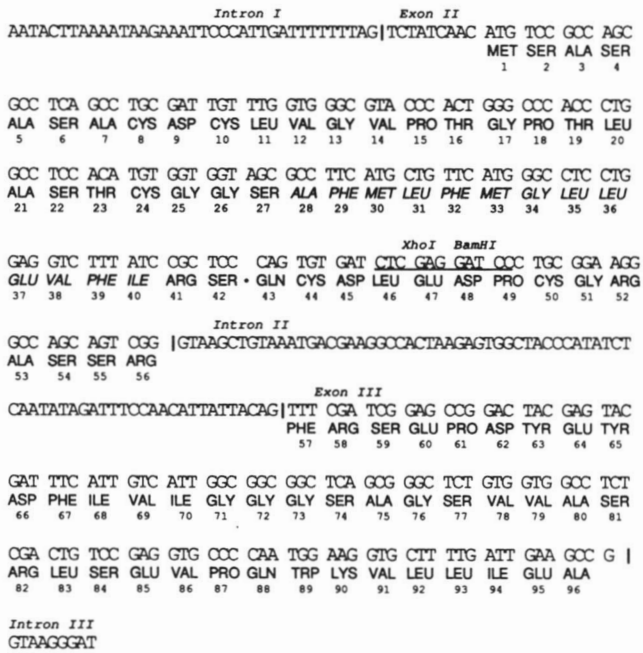

FIGURE 7.—DNA sequence of exon II, intron II, and exon III. Vertical bars denote exon-intron boundaries. The inferred amino acid sequence is given below the DNA sequence. Amino acid sequences in italics (residues 28–40) denotes a highly hydrophobic region. The dot between residues 42 and 43 denotes a putative signal peptide cleavage site. However, see DISCUSSION.

are poorly conserved (P. KRASNEY and D. CAVENER, unpublished data).

To confirm that no *Gld* exons exist between exons I and II, a primer extension experiment was performed using a primer corresponding to exon II (at the *Bam*HI site). If no other exonic sequences lie between exons I and II, this experiment should yield a 490-nt fragment. Three primer extension products were observed: 180, 185 and 485 nt (data not shown). The latter is very close to the predicted fragment. The smaller primer extension products are most likely the result of the primer binding to partially complementary sites in exon I.

The Pustell-IBI codon bias method was used to search for a putative protein coding region in exons I, II and III (Figure 9). It is almost certain that exon I is entirely untranslated since it does not contain any start codons and does not exhibit a codon bias typical of Drosophila genes. A 96 amino acid open reading frame was identified which begins 10 nt from the 5′ end of exon II and continues through exon III (Figure 7). The putative start codon is flanked by sequences which are similar to the Drosophila consensus sequence (C/A A A A/C A U G) for translation initiation sites (CAVENER 1987b). That this is the translation start site and a functional coding sequence is strongly supported by phylogenetic comparisons of these sequences among three divergent *Drosophila* species (P. KRASNEY and D. CAVENER, unpublished data).

Functional delimitation of *Gld* via *P* element-mediated transformation: The *Gld* gene was isolated by the method of chromosome walking (CAVENER *et al.* 1986a) and identified by the localization of three independent *Gld* mutations (Figure 1) which pinpoint its chromosomal location (CAVENER, OTTESON and KAUFMAN 1986). In order to provide a functional proof for the existence of the *Gld* gene within the cloned genomic DNA described above, we used the method of *P* element-mediated germline transformation (RUBIN and SPRADLING 1983). A 17.5-kb *Kpn*I fragment (pieced together from two lambda genomic clones) was inserted into the pW5 *P* element transformation vector (KLEMENZ, WEBER and GEHRING 1987) to generate pWG67 (Figure 10). Two independent germline transformants carrying pWG67 were obtained (T-7.1 and T-14.3) which expressed the visible marker (*white* gene) of the transformation vector. Genomic Southern analysis indicated that both transformants contain a single copy of the transforming DNA integrated into the genome (Figure 11). Flies from these two transformed lines were crossed into a *Gld* mutant background in order to determine if the 17.5-kb *Kpn*I fragment could rescue the lethal *Gld* mutant phenotype. Both transformants were found to

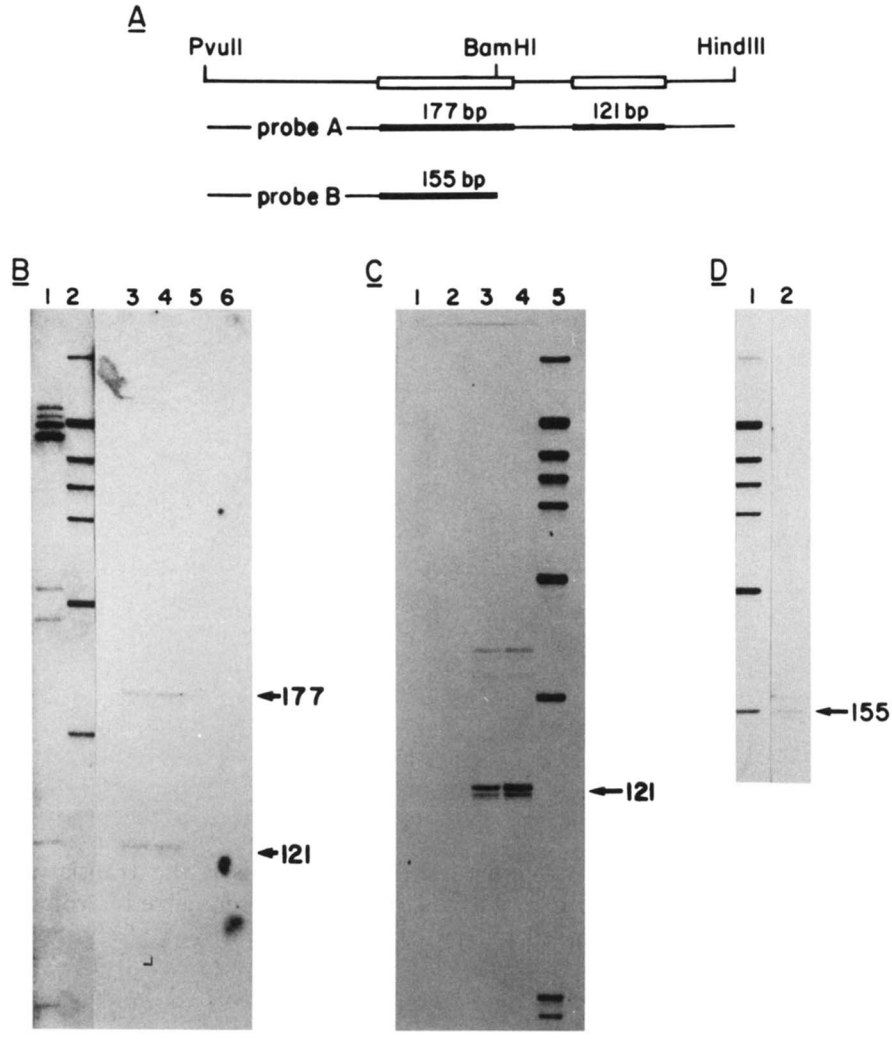

FIGURE 8.—RNase protection mapping of exon II, intron II, and exon III. (A) The final derived map of exon II and exon III (open boxes) and intron II (solid line between the two exons). The complementary cRNA probes used to determine the size and boundaries of exon II and exon III are below the map. The thick portion of each probe map represents the protected fragment after RNase digestion. (B) Probe A. Lanes 1 and 2, molecular weight markers; Protecting RNAs: lane 3, male; lane 4, pharate adult; lane 5, female; and lane 6, yeast. (C) Probe A, partially linearized at the *Bam*HI site. This yields two cRNA probes: a full length *Pvu*II-*Hin*dIII probe and a *Bam*HI-*Hin*dIII probe. Protecting RNAs: lane 1, yeast; lane 2, female; lane 3, pharate adult; and lane 4, adult male. Molecular weight markers in lane 5. In addition, a 27-nt fragment was also observed in lanes 3 and 4 (not shown) which corresponds to a region in exon II from the *Bam*HI site to the 3′ end. (D) Probe B. Molecular weight markers—1; Protecting RNAs: 2—adult male.

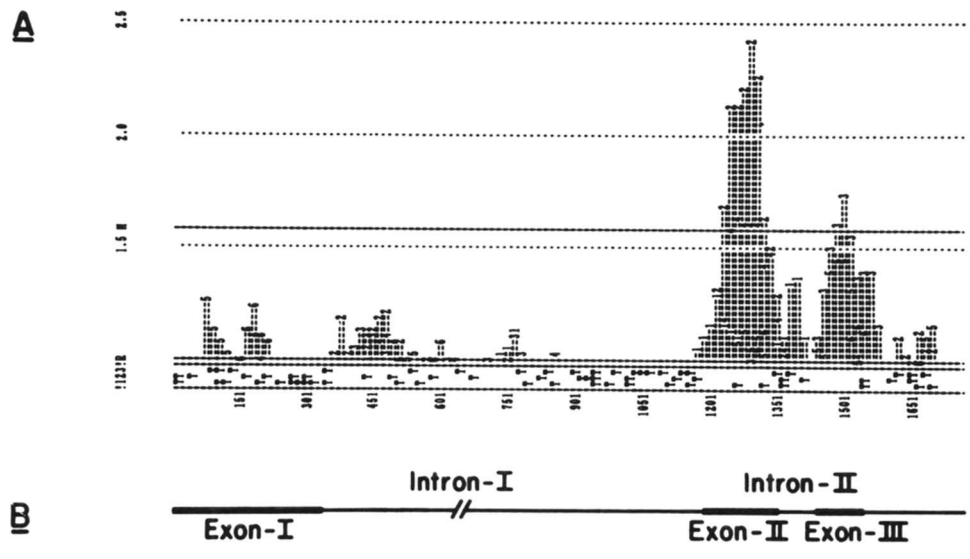

FIGURE 9.—Codon bias analysis of exons I, II and III. The DNA sequence file of exon I and the first 307 nt of intron I was fused to a sequence file containing the last 529 nt of intron I, exon II, intron II, exon III and the first 215 nt of intron III. This composite sequence was analyzed using the Pustell-IBI Protein Coding Region Locator program on a Compaq 286 computer. The vertical dashed lines represent the value of the *C*-statistic calculated for successive 40 nt steps. At the top of each line is a number (1–3) indicating the reading frame. Values which extend above the line labelled M indicate regions which display significant bias in codon usage when compared with a Drosophila codon bias table. This analysis identifies two putative coding regions located in exons II and III. The "T"s below the *C*-statistic values denote the positions of termination codons in reading frames 1, 2 and 3.

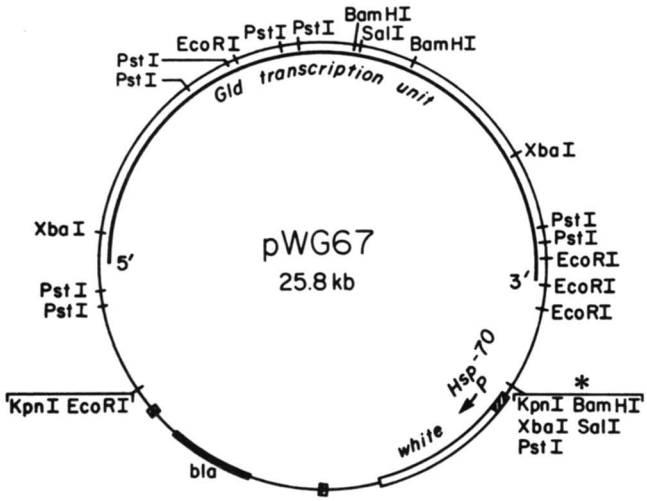

FIGURE 10.—Map of the *Gld* transforming recombinant clone. pWG67 is composed of a 17.5-kb *Kpn*I fragment containing the *Gld* gene (see Figure 1) in the pW5 *P* element-transformation vector (KLEMENZ, WEBER and GEHRING 1987). The vector contains a Hsp-70 promoter (slashed box) fused to the coding region of the *white* gene (open box) as the selectable marker in Drosophila, the β-lactamase gene (*bla*, solid box) for ampicillin resistance, and a *P* element with the 31-bp terminal inverted repeats (boxed arrowheads). A complete restriction map is given only for *Eco*RI, *Pst*I, *Bam*HI, *Sal*I, *Xba*I, and *Kpn*I. (*) Additional restriction sites (not shown) are present in the polylinker site.

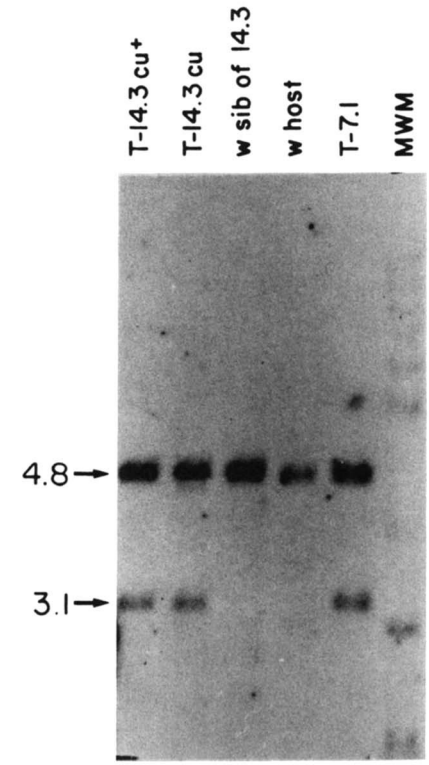

FIGURE 11.—Genomic Southern analysis of the T-7.3 and T-14.1 *Gld* transformants. *Xba*I/*Eco*RI double digests of 15 μg of genomic DNA for each sample. Blots were hybridized with a ^{32}P nick translated probe of pCG4 (coordinates 147.4–151.6 of Figure 1) containing exon I. It should be noted that the *Eco*RI site in pCG4 corresponds to a vector/insert junction restriction site of lambda clone E14b (CAVENER *et al.* 1986b) and not to an actually restriction site in the *Gld* gene. The relevant *Eco*RI site in the Drosophila genomic DNA is 600 bp upstream of the endpoint of pCG4. The 4.8-kb fragment observed in each lane corresponds to the *Xba*I/*Eco*RI fragment of the endogeneous *Gld* gene (coordinates 147.4–152.2 of Figure 1). The 3.1-kb fragment corresponds to a unique *Xba*I/*Eco*RI fragment of pWG67 (see Figure 10). The 3.1-kb band is only observed in the three lanes containing DNA from the w^+ transformants. Note that the intensity of the 3.1-kb band is less than half the intensity of the 4.8 kb. This result was expected since the pWG67 inserts were heterozygous in the sampled flies. In addition, the relative intensity level of the 3.1-kb band indicated that both the transformants contained single pWG67 inserts. The assumption that both transformants contained single inserts was confirmed by additional Southern hybridization experiments (data not shown).

completely rescue *Gld* mutants from their non-eclosion lethal phenotype. Moreover, these transformants exhibit the normal temporal pattern of *Gld* expression: low expression in feeding third instars, high expression in wandering third instars (*i.e.*, immediately before pupariation), high expression during metamorphosis and in adult males and very low expression in adult females (Table 1). The quantitative levels of GLD expression are quite similar between the host and transformant lines, although GLD activity in T-14.3 adult males is significantly lower than what we have observed for a variety of wild-type strains. The latter effect is probably due to the influence of the local genomic environment of this transformant.

DISCUSSION

Gene structures in Drosophila have often been dichotomized between small genes (*ca.* 1–7 kb) which encode enzymes or other nonregulatory proteins and large genes (*ca.* 50–100 kb) which encode proteins which function to regulate development (LEWIN 1987). However, the *Gld* gene (*ca.* 18 kb) joins a growing list of intermediate sized genes encoding enzymes (*Dunce*—CHEN, DENOME and DAVIS 1986; *Gart*—HENIKOFF *et al.* 1986; *Ace*—HALL and SPIERER 1986) which obviate this dichotomy. One structural feature which is generally shared by Drosophila genes of all lengths is the presence of a small 5′ exon usually containing the untranslated leader sequence and a very small portion of the coding region or, in few cases such as *Gld*, containing only the leader sequence. The first exon is then followed by what is typically the largest intron of the gene (*e.g.*, *yellow*—CHIA *et al.* 1986; large subunit of RNA polII—BIGGS, SEARLES and GREENLEAF 1985; *Gart*—HENIKOFF *et al.* 1986; α1, α2 and α4 tubulin genes—THEURKAUF *et al.* 1986). In the case of *Gld* the first intron is unusually large (8 kb). We speculate that these structural features may be the result of independent origins of the regulatory and coding regions consistent with the exon shuffling model of gene evolution (GILBERT 1978) or that such gene structures are the result of random acquisition of regulatory elements and coding

TABLE 1

GLD enzyme activities[a] for pWG67 transformants[b]

Stage	Transformants		Oregon-R
	T-7.1	T-14.3	
Feeding 3rd instar	7.1 (2.0)	5.1 (0.7)	5.2 (0.6)
Wandering 3rd instar	15.3 (6.2)	25.2 (3.0)	30.8 (1.2)
Prepupae/early pupae	51.7 (20.2)	22.2 (0.7)	35.0 (9.6)
Pharate adults	64.0 (18.8)	74.6 (3.1)	59.5 (21.0)
Adult females	6.9 (0.7)	8.3 (2.3)	11.3 (1.2)
Adult males	128.0 (1.5)	40.6 (1.4)	94.6 (1.9)

[a] Micromoles of DCIP reduced min^{-1} $individual^{-1}$. Each value represents the mean of 3–6 replicates. Standard errors are given in parentheses.

[b] Both transformants are in *Gld* null mutant genetic backgrounds. The T-7.1 strain is homozygous with respect to the pWG67 insert, whereas T-14.3 is hemizygous.

sequences along a contiguous linear sequence. The latter idea is a simple extension of the model proposed by SENAPATHY (1986) for the evolution of coding exons.

Upstream of the *Gld* transcription start site is a somewhat unusual TATA sequence (−31, TTTAAAAA) similar to that found for the Drosophila *dopa decarboxylase* gene (HIRSCH, MORGAN and SCHOLNICK 1986). Two interesting sequence elements found upstream of the TATA box are: (1) a 13-bp palindrome (at −73) separated by a single base pair at the axis of dyad symmetry and (2) four copies of a 7 bp dispersed repeat (at −248, −154, −144 and −106) based upon the sequence motif TAGACCA. A search in the promoter regions of a number of Drosophila genes in the Genbank data base and recent publications failed to detect these particular sequences in other known and putative promoters. In addition to the sequence elements immediately upstream of the start site of transcription, a 72-bp tetranucleotide tandem repeat element was found in the middle of intron I (CAVENER *et al.* 1988). This sequence element, named the YYRR box, is conserved in the *Gld* gene of three divergent Drosophila species indicating that it may serve some function. The requirement of these sequence elements for *Gld* expression is currently under investigation using the techniques of *in vitro* mutagenesis and P-element mediated transformation. The transformation experiments reported herein indicate that an 17.5-kb *Kpn*I genomic fragment, which includes 2.3 kb of sequence to the 5′ side of the *Gld* transcription start site, is sufficient for normal quantitative and qualitative expression.

Another unusual feature of the *Gld* gene is the presence of a large (344 nt) untranslated leader sequence. Although long leader sequences were once thought to be unique to heat shock genes (SOUTHGATE, AYME and VOELLMY 1983), the number of other *Drosophila* genes which have been reported to contain long leader sequences (*i.e.*, >200 nt) has dramatically increased in the past few years (*e.g.*, *yellow*—CHIA *et al.* 1986; *Antennapedia*—LAUGHON *et al.* 1986 and STROEHER, JORGENSEN and GARBER 1986; *Ace*—HALL and SPIERER 1986; *Kruppel*—ROSENBERG *et al.* 1986; *Notch*—WHARTON *et al.* 1985 and KIDD, KELLY and YOUNG 1986; large subunit of RNA polII—BIGGS, SEARLES and GREENLEAF 1985; *Ultrabithorax*—WILDE and AKAM 1987; *Dint-1*—RIJSEWIJK *et al.* 1987). The discovery of long leader sequences among these genes has raised the question of translational control particularly since virtually all such leaders contain multiple short upstream open reading frames (uORFs). The *Gld* leader sequence is thus somewhat unique among long leaders in being devoid of uORFs.

The GLD enzyme is secreted into the molting fluid of pupae and into the seminal fluid of adult males. Thus we expected to find a hydrophobic signal sequence at the inferred amino terminus of GLD. Although some hydrophobic residues are located among the first 27 amino acids, a highly hydrophobic region is not found until residues 28 through 40. However, analysis of this putative signal sequence (Figure 7) using the weight-matrix procedure of HEIJNE (1986) indicated that it did not conform particularly well to the eukaryotic signal sequence consensus (analysis not shown). In addition, no other region in the first fifty residues of the inferred GLD preprotein was found to conform significantly better. Obviously, the location of the signal peptide and the cleavage site must await direct sequence analysis of the mature GLD protein. Two other unusual features of the inferred amino terminus are the serine-alanine triplet repeat (residues 2–7) and the presence of five cysteine residues.

We thank SUSAN SCHLITZ for technical assistance and ROBERT THOMPSON for advice on the primer extension experiments. This work was supported by grants from the National Institutes of Health and the National Science Foundation to D.C.

LITERATURE CITED

BIGGS, J., L. L. SEARLES and A. L. GREENLEAF, 1985 Structure of the eukaryotic transcription apparatus: features of the gene for the largest subunit of *Drosophila* RNA polymerase II. Cell **42:** 611–621.

CAVENER, D. R., 1987a Combinatorial control of structural genes in Drosophila: Solutions that work for the organism. BioEssays **7:** 103–107.

CAVENER, D. R., 1987b Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. Nucleic Acids Res. **15:** 1353–1361.

CAVENER, D. R., and R. J. MACINTYRE, 1983 Biphasic expression and function of glucose dehydrogenase in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **80:** 6286–6288.

CAVENER, D. R., D. OTTESON and T. C. KAUFMAN, 1986 A rehabilitation of the genetic map of the 84 B-D region in *Drosophila melanogaster*. Genetics **114:** 111–123.

CAVENER, D., G. CORBETT, D. COX and R. WHETTEN, 1986a Isolation of the eclosion gene cluster and the developmental

expression of the *Gld* gene in *Drosophila melanogaster*. EMBO J. **5:** 2939–2948.

CAVENER, D. R., M. MURTHA, C. SCHONBAUM and D. HOLLAR, 1986b Cell autonomous and hormonal control of sex-limited gene expression in *Drosophila*. UCLA Symp. Mol. Cell. Biol. **49:** 453–462.

CAVENER, D., Y, FENG, B, FOSTER, P. KRASNEY, M. MURTHA, C. SCHONBAUM and X. XIAO, 1988 The YYRR box: a conserved dipyrimidine-dipurine sequence element in *Drosophila* and other eukaryotes. Nucleic Acids Res. **16:** 3375–3390.

CHEN, C. N., S. DENOME and R. L. DAVIS, 1986 Molecular analysis of cDNA clones and the corresponding genomic coding sequences of the *Drosophila* dunce gene, the structural gene for cAMP phosphodiesterase. Proc. Natl. Acad. Sci. USA **83:** 9313–9317.

CHEN, E., and P. H. SEEBURG, 1985 Supercoil sequencing. DNA **4:** 165–170.

CHERBAS, L., R. A. SCHULZ, M. M. D. KOEHLER, C. SAVAKIS and P. CHERBAS, 1986 Structure of the Eip28/29 gene, an ecdysone-inducible gene from Drosophila. J. Mol. Biol. **189:** 617–631.

CHIA, W., G. HOWES, M. MARTIN, Y. B. MENG, K. MOSES and S. TSUBOTA, 1986 Molecular analysis of the yellow locus of *Drosophila*. EMBO J. **5:** 3597–3605.

DENTE, L., G. CESARENI and R. CORTESE, 1983 pEMBL: a new family of single stranded plasmids. Nucleic Acids Res. **11:** 1645–1655.

FISCHER, J. A., and T. MANIATIS, 1985 Structure and transcription of the *Drosophila mulleri* alcohol dehydrogenase genes. Nucleic Acids Res. **13:** 6899–6916.

GILBERT, W., 1978 Why genes in pieces? Nature **271:** 501.

HALL, L. M. C., and P. SPIERER, 1986 The Ace locus of *Drosophila melanogaster*; structural gene for the acetylcholineasterase with an unusual 5′ leader. EMBO J. **5:** 2949–2954.

HEIJNE, G., 1986 A new method for predicting signal sequence cleavage sites. Nucleic Acids. Res. **14:** 4683–4690.

HENIKOFF, S., 1984 Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28:** 351–359.

HENIKOFF, S., 1987 Unidirectional digestion with exonuclease III in DNA sequence analysis. Methods Enzymol. **155:** 156–165.

HENIKOFF, S., M. A. KEENE, K. FECHTEL and J. W. FRISTROM, 1986 Gene within a gene: nested *Drosophila* genes encode unrelated proteins on opposite DNA strands. Cell **44:** 33–42.

HIRSCH, J., B. A. MORGAN and S. B. SCHOLNICK, 1986 Delimiting regulatory sequences of the *Drosophila melanogaster* Ddc gene. Mol. Cell. Biol. **66:** 4548–4557.

KIDD, S., M. KELLEY and M. W. YOUNG, 1986 Sequence of the Notch locus of *Drosophila melanogaster*: Relationship of the encoded protein to mammalian clotting and growth factors. Mol. Cell. Biol. **6:** 3094–3108.

KLEMENZ, R., U. WEBER and W. J. GEHRING, 1987 The white gene as a marker for gene transfer in *Drosophila*. Nucleic Acids Res. **15:** 3947–3959.

LAUGHON, A., A. M. BOULET, J. R. BERMINGHAM, R. A. LAYMON and M. P. SCOTT, 1986 Structure of transcripts from the homeotic *Antennapedia* gene of *Drosophila melanogaster*: two promoters control the major protein-coding region. Mol. Cell. Biol. **6:** 4676–4689.

LEWIN, B., 1987 *Genes*, Ed. 3, p. 693. John Wiley & Sons, New York.

MANIATIS, T., E. F. FRITSCH and J. SAMBROOK, 1982 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

MOUNT, S., 1982 A catalogue of splice junction sequences. Nucleic Acids Res. **10:** 459–472.

RIJSEWIJK, F., M. SCHUERMANN, E. WAGENAAR, P. PARREN, D. WEIGEL and R. NUSSE, 1987 The *Drosophila* homolog of the mouse mammary oncogene *int-1* is identical to the segment polarity gene *wingless*. Cell **50:** 649–657.

ROSENBERG, U. B., C. SCHRODER, A. PREISS, A. KIENLIN, S. COTE, I. RIEDE and H. JACKLE, 1986 Structural homology of the product of the *Drosophila Kruppel* gene with *Xenopus* transription factor IIIA. Nature **319:** 336–339.

RUBIN, G. M., and A. C. SPRADLING, 1983 Genetic transformation of *Drosophila* with transposable element vectors. Science **218:** 348–353.

SANGER, F., and A. F. COULSON, 1975 A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. **94:** 444–448.

SENAPATHY, P., 1986 Origin of eukaryotic introns: a hypothesis, based on codon distribution statistics in genes, and its implications. Proc. Natl. Acad. Sci. USA **83:** 2133–2137.

SNYDER, M., M. HUNKAPILLER, D. YUEN, D. SILVERT, J. FRISTROM and N. DAVIDSON, 1982 Cuticle protein genes of Drosophila: structure, organization and evolution of four clustered genes. Cell **29:** 1027–1040.

SOUTHGATE, R., A. AYME and R. VOELLMY, 1983 Nucleotide sequence analysis of the *Drosophila* small heat shock gene cluster at locus 67B. J. Mol. Biol. **165:** 35–57.

SPRADLING, A. C., and G. M. RUBIN, 1983 Transposition of cloned P elements into *Drosophila* germ line chromosomes. Science **218:** 341–347.

STROEHER, V. L., E. M. JORGENSEN and R. L. GARBER, 1986 Multiple transcripts from the *Antennapedia* gene of *Drosophila melanogaster*. Mol. Cell. Biol. **6:** 4667–4675.

THEURKAUF, W. E., H. BAUM, J. BO and P. C. WENSINK, 1986 Tissue-specific and constitutive α-tubulin genes of *Drosophila melanogaster* code for structurally distinct proteins. Proc. Natl. Acad. Sci. USA **83:** 8477–8481.

WHARTON, K. A., K. M. JOHANSEN, T. XU and S. ARTAVANIS-TSAKONAS, 1985 Nucleotide sequence from the neurogenic locus *Notch* implies a gene product that shares homology with proteins containing EGF-like repeats. Cell **43:** 567–581.

WILDE, C. D., and M. AKAM, 1987 Conserved sequence elements in the 5′ region of the Ultrabithorax transcription unit. EMBO J. **6:** 1393–1402.

ZINN, K., D. DIMAIO and T. MANIATIS, 1983 Identification of two distinct regulatory regions adjacent to the human β-interferon gene. Cell **34:** 865–879.

Communicating editor: V. G. FINNERTY