Volume 9 Number 23 1981 Nucleic Acids Research

The sequence of the Drosophila melanogaster gene for yolk protein 1

Mien-Chie Hung and Pieter C.Wensink

Department of Biochemistry and Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02254, USA

Received 21 September 1981

ABSTRACT

We have determined the complete nucleotide sequence of the hormonally regulated yolk protein 1 (YP1) gene of *Drosophila melanogaster*. We have also determined the sequence location of the 5' and 3' ends of both the mature mRNA and the gene's only intron. The YP1 gene contains sequences similar to those found in many other eukaryotic genes. Among these sequences are the Hogness-Goldberg box, the capping site, the ribosome binding site and the polyadenylation signal sequence, all perhaps involved in transcriptional or translational control. Also among these sequences are the consensus splice sequences. They occur at each end of the 76 nucleotide intron. One distinctive secondary structure likely to occur in either the DNA or RNA and which might be involved in the regulation of transcription or translation was also found in the YP1 gene sequence. We show the protein sequence predicted by the DNA sequence and the RNA mapping results.

INTRODUCTION

Three yolk proteins (YP1, YP2 and YP3) have been identified in *D. melanogaster* (1). These proteins are coordinately synthesized by the female fat body in response to a steroid hormone, ecdysone (2). We have recently isolated genes coding for these proteins (3) and demonstrated that each protein is encoded by a different, single-copy gene. Since we have also established that the concentration of yolk protein mRNAs in the fat body increases in parallel with the rate of the fat body's synthesis of these proteins (4), we presume that the hormone controls the rate of protein synthesis by controlling the rate at which these three single-copy genes are transcribed. A transcription map of the genes (5), indicates that this control is exerted at a minimum of three transcription initiation sites, one site for each of the genes. As part of our investigation of the mechanism which allows the hormone to exert this control, we present in this paper the entire DNA sequence of the YP1 gene and establish the location of the 3' and 5' ends of both the mRNA and the gene's single intron.

MATERIALS AND METHODS

Plasmid DNAs and *D. melanogaster* RNAs were isolated as described by Barnett *et al.* (3).

DNA fragments were radiolabeled at their 5' end with T4 polynucleotide kinase (Bethesda Research Laboratories) by the method of Maxam and Gilbert (6) or at their 3' end with avian myeloblastosis virus reverse transcriptase (obtained from J. Beard) by the method of Goodman (7).

Restriction mapping and DNA sequencing were performed according to Smith and Birnstiel (8) and to Maxam and Gilbert (6), respectively. The G, G+A, C+T and C modification reactions (6) were used for the sequencing.

The nuclease protection studies were carried out according to the general procedure of Sollner-Webb and Reeder (9) using sequencing ladders to determine the end points of protected fragments that were end-labelled at either their 3' or 5' end. Because Maxam and Gilbert sequencing ladders were used as size standards in these experiments, correction factors had to be applied. The sequencing reactions eliminate the modified nucleotide. Thus fragments generated by the sequencing reactions are one nucleotide shorter than the corresponding fragments generated by S1-digested product for 3' end-labelled DNA fragments. For 5' end-labelled DNA fragments, as shown by Sollner-Webb and Reeder (9), the sequencing cleavage products have an apparent electrophoretic mobility 1.5 nucleotides faster than the corresponding fragments generated by S1-digested products because the sequencing cleavage products are one nucleoside shorter than the corresponding SI-digested product.

The regions of dyad symmetry were identified using the computer program of Queen and Korn (10) in the Stanford molgen project version. When this program was used the relevant parameters of Queen and Korn were set as: MIN MATCH = 5; MIN RATIO = 0.7; LOOP LENGTH = 3; LOOP DIST = 50; DUBIOUS = 0.

RESULTS AND DISCUSSION

DNA Sequence of the YP1 Gene

The YP1 gene sequence was determined by examining genomic DNA cloned (3) in the plasmid pBR322. The sequencing strategy and the map of relevant restriction endonuclease sites are shown in Figure 1. As this figure indicates, approximately 80% of the gene sequence was confirmed by sequencing both complementary strands. The remainder of the sequence was unambiguously determined by the sequence of one or the other strand. Since our restriction map was of low resolution, there is the possibility that any one of the mapped

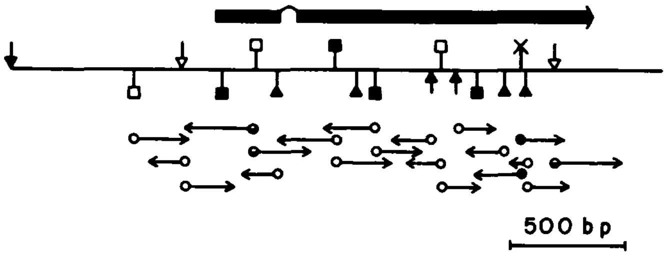

Figure 1: Strategy for sequencing the YP1 gene. The restriction map is drawn with Hind III (↓), EcoR I (↑), BamH I(□), BstE II(■), Bgl II(□), Ava II(■), Xho I(✗), Pst I(▲), and Hinf I(▽) sites shown. The only Hinf I sites shown are those used in the experiments. All sites of the other enzymes are shown. The arrows under the restriction map indicate the direction and the extent of DNA sequence determination. The circles at the end of the arrows indicate the strand labeled and sequenced: 0, 5' end-labeled by kinase; ●, 3'end-labeled by reverse transcriptase; and ◐, both 5' and 3' ends were labeled and sequenced. The heavy lines above the restriction map represent the DNA region complementary to YP1 mRNA (5). Caret symbol indicates the intron and the arrowhead indicates the 3' end of mRNA.

sites might be a doublet of two closely spaced, identical restriction sites. To assure that no part of the gene sequence was lost because of this possibility, we labeled additional sites and determined the sequence across restriction sites used in the sequencing procedure. The single exception to this general practice was the BamH I site closest to the 3' end of the gene. We demonstrated that this site was not a doublet by digesting the entire plasmid with BamH I, labeling the ends and then electrophoresing the product on sequencing gels that would detect fragments from 2 to at least 2000 nucleotides in length. No fragment was observed between the unincorporated γ-^{32}P-ATP and the smallest fragment predicted by the restriction map (800 bp, see Fig. 1). This observation was confirmed by electrophoresis of a partial BamH I digest of a BstE II/Xho I fragment (see Fig. 1) labeled at its BstE II end. No deletion larger than 5 or 6 nucleotides was found in the partial digest. This indicates that the BamH I site was not a doublet. We conclude that the complete genomic copy of the YP1 gene was determined. It is shown in Figure 2.

We observed that the second cytosine of every BstN I site (5'-CC$\genfrac{}{}{0pt}{}{A}{T}$GG-3') was always missing in the sequencing gel. This is likely to be due to methylation of this residue when the cloned DNA was grown in _E. coli_ (11). All of the BstN I sites predicted by the sequence were confirmed by restriction mapping, and in most cases (see Fig. 1) were further confirmed by

```
...AGATCTATATTTTATGCATTTATTTGATCAAATCCGGTGCACAACTACAATGTTGCAATCAGCGGAACTACAAAGTG -271
ATTACAAATTAAAATAATCAGGCGGCAGCAGGTGCTGCTAAGTCATCAGTGGGGTCAGCTATAGGTAGGCCCCGTGTCTA -191
TTTTGTATGTATACAATTTATTCCGCTATCGATAGCATATACACTCATCCGATTCCTTAGGCACCCGAAAACCCTTACTC -111
AGCACAAGTGACCGATTAAGGCCTGAGCCAGCGAAAAGCAAGTCGGAAAATGGGAAATCGCTCAGCGTAAATTGTGGTAT  -31
ATAAACCACCATCGTTGGATTTGGAAGGCCAGTTCAACTCACTCAGTGTTGAAGTCGCATCCGCAGGACCAAATCCCAAA   50
TCCGAACCATGAACCCCATGAGAGTGCTGAGCCTTCTGGCTTGCTTGGCGGTCGCCGCCTTGGCCAAGCCCAATGGCCGT  130
ATGGACAACTCCGTCAACCAGGCATTGAAGCCGTCGCAGTGGCTCTCCGGATCCCAGCTGGAGGCCATTCCCGCCCTCGA  210
CGATTTCACCATTGAGCGTCTGGAGAACATGAACCTGGAGCGTGGCGCCGAGCTGCTGCAGCAAGTCTGTGAGTAATCCT  290
AGATGCAGATAAAAAAAAAAAAAAAAACATCGAATATTCTATGGAATATATATATCCTTTGTAGACCACCTGTCGCAGAT  370
CCACCACAACGTTGAGCCCAACTATGTGCCCAGCGGCATCCAGGTCTATGTGCCCAAGCCCAATGGTGACAAGACCGTTG  450
CTCCCCTGAACGAGATGATCCAGCGCCTGAAGCAGAAGCAGAACTTTGGTGAGGATGAGGTGACCATCATTGTGACCGGA  530
CTGCCCCAGACCAGCGAGACCGTGAAGAAGGCGACCAGGAAGCTGGTTCAGGCTTACATGCAGCGCTACAATCTGCAGCA  610
GCAGCGCCAGCACGGCAAGAACGGCAACCAGGACTACCAGGATCAGAGCAACGAACAGAGGAAGAACCAGAGGACCAGCA  690
GCGAGGAGGACTACAGCGAGGAGGTTAAGAACGCCAAGACCCAAAGCGGCGACATCATTGTGATCGATTTGGGCTCCAAG  770
CTGAACACCTATGAGCGTTATGCCATGCTCGACATTGAGAAGACCGGCGCCAAGATCGGCAAGTGGATCGTCCAGATGGT  850
CAACGAGTTGGACATGCCCTTCGATACCATTCACCTGATTGGCCAGAATGTGGGTGCCCATGTTGCCGGTGCCGCTGCCC  930
AGGAATTCACCCGTCTCACCGGACACAAGCTGCGCCGTGTCACCGGTCTGGATCCCTCCAAGATCGTGGCCAAGAGCAAG 1010
AACACCCTGACCGGTCTGGCTCGCGGTGATGCTGAATTCGTTGACGCCATCCACACCTCGGTCTACGGCATGGGCACCCC 1090
CATCCGCTCCGGTGATGTTGACTTCTATCCCAATGGACCTGCCGCCGGTGTTCCCGGAGCCAGCAACGTGGTGGAGGCCG 1170
CCATGCGTGCCACCCGCTACTTCGCCGAGTCCGTGCGTCCCGGAAACGAGAGGAGCTTCCCCGCCGTGCCAGCCAACTCC 1250
CTGCAGCAGTACAAGCAGAACGATGGATTCGGCAAGCGTGCCTACATGGGCATCGATACCGCTCACGATCTCGAGGGTGA 1330
CTACATTCTGCAGGTGAACCCCAAGTCTCCTTTCGGCCGCAACGCACCCGCCCAGAAGCAGAGCAGCTACCACGGTGTCC 1410
ACCAGGCGTGGAACACCAACCAGGACAGCAAGGACTACCAGTAAGGATGAGTCTGCTTACTCTGGACACCTGGAATGGCA 1490
ACTACCAAACAACCACCCAACCACACAAACACTGTAGTCCCTAAGTTGAACCCATATTGGCCCTTTTCTTGAGATTACCT 1570
AAACATTTAACGAGCACATCGCGAAATTCAGCAAATAAACGCTCGATAAAGAGCTTAAAAATATCTATTTTGTTTATCTT 1650
AAATCATTTAGGAACTATAATAGTCTAATAGATCATCCCAAAAAAAAGGGAACAAAATCAAAAGTAAATATCGTAGTTTG 1730
GTTTTGTAAACTTAGATTTATTTTATTGTTGTCGGTGTTTTTGTGG.....
```

Figure 2: Complete nucleotide sequence of the YP1 gene. The nucleotide sequence shown is the non-coding strand and corresponds to the mRNA sequence. It is written from left to right and top to bottom, in the 5' to 3' direction. Some of the relevant restriction sites are indicated along the sequence by the symbols: that is BamH I, ▽ ; for BstE II, ▼ ; and for Hinf I, ◇ . In 5' to 3' order, Hogness-Goldberg box, capping site, ribosomal binding site, first ATG codon, termination codon (TAA), and polyA recognition site are all outlined by heavy lines. The arrows indicate the 5' and 3' ends of the YP1 gene and the intron. The intron region is marked by underlining.

sequencing both of the complementary strands.

Location of the 5' and 3' ends of the mRNA

A map of the YP1 mRNA complementarity to the YP1 gene had been established by single-strand specific nuclease digestion of DNA:RNA duplexes (5). This transcript map indicated that the YP1 gene contained, from its 5' to its 3' end, a 290 nucleotide exon, a 70 nucleotide intron and a 1300 nucleotide exon. In this paper we describe similar experiments which differ by using sequencing gel ladders to determine the precise gene sequence location of the 5' and 3' ends of both the YP1 mRNA and the intron.

The strategy used to localize the ends is summarized in Figure 3a. To map the 5' end of the YP1 mRNA, the 5' end-labeled fragment A (Fig. 3a) was denatured and hybridized to adult polyA$^+$ RNA under conditions which favored DNA:RNA hybridization rather than DNA:DNA hybridization (12). Following the hybridization reaction, single-stranded nucleic acids were digested with S1 nuclease. The labeled, S1-resistant duplexes were denatured and electrophoresed on a 6% polyacrylamide DNA sequencing gel in parallel with the 4 sequencing reaction products from fragment A (Fig. 3b). Three S1 resistant fragments were observed whose termini were 182, 184 and 185 nucleotides from the end of fragment A. As shown in Figure 3c, these termini correspond to positions -2, -1 and +2 of the gene sequence (Fig. 2). Since the S1 nuclease can nibble at the ends of RNA:DNA duplexes (9), the 5' end of the mRNA could be at any one of these three positions or at a nearby position. It is also possible that there may be several mRNAs with slightly different end points. The sequence in the area of the 5' end indicates which nucleotide is most likely to be the 5' end. The terminus of the largest S1 resistant fragment (position -2 in Fig. 2) begins an 8 nucleotide sequence (5'-CCAGTTCA-3') that is similar to the consensus sequence for the eukaryotic mRNA capping site (5'-PyCATTCPu-3') (13,14). The 5' ends of most eukaryotic mRNAs are located at an adenosine that is the third nucleotide of this capping site (13,14). In addition, one experiment described in the literature indicates that transcription is initiated within the capping site (15). Reasoning by analogy, we suggest that the 5' end of the YP1 mRNA corresponds to the adenosine (position +1 in Fig. 2) that is bracketed by the termini of S1-resistant fragments.

Analogy to another consensus sequence also suggests that the 5' end of the mRNA is near this adenosine. A sequence identical to the Hogness-Goldberg box (5'-TATATAAA-3') occurs 26 nucleotides upstream from the adenosine at the +1 position. This box is found 20-30 nucleotides upstream from the mRNA start

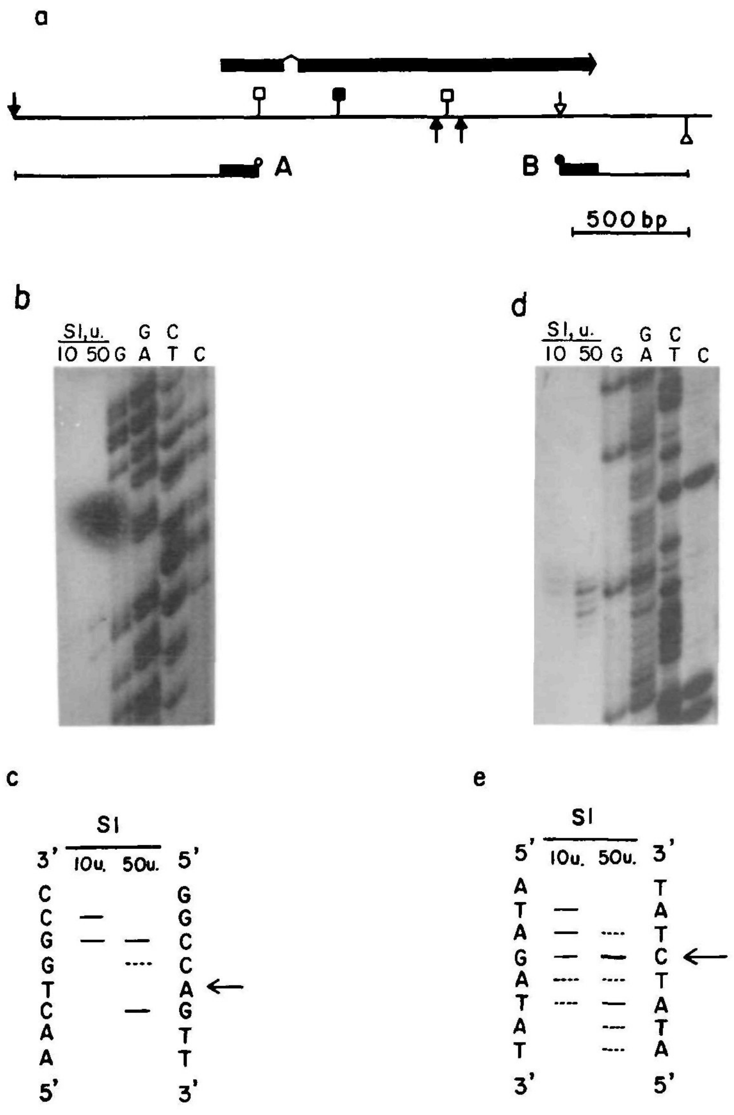

Figure 3: Location of 5' and 3' ends of YP1 gene.
(a) A simplified restriction map is shown. The symbols are the same as in Figure 1 except that a Cfo I site (⊥) is present. A and B indicate the restriction fragments used to map the 5' and 3' ends, respectively, of the mRNA. The DNA protected from S1 digestion is shown by heavy lines on the restriction fragments. Open and closed circles represent labels at the 5' and 3' ends, respectively.
(b) The S1-resistant material from fragment A was electrophoresed on a 6% polyacrylamide DNA sequencing gel. The amount of S1 nuclease used in 200 µl of reaction buffer is shown above the lanes. A DNA sequencing ladder prepared

from the same fragment was used a size marker.
(c) A schematic representation of (b) is shown. The sequence read from (b) is shown to the left. Its complementary strand (the non-transcribed strand which corresponds to the mRNA sequence) shown to the right. The sequencing standards are displaced by 1.5 nucleotides to account for the difference in their migration relative to the S1 nuclease-generated fragments (see Materials and Methods). The arrow indicates the proposed 5' end of YP1 gene.
(d) The S1-resistant material from fragment B is shown. The conditions of electrophoresis and digestion as well as the lane pattern are the same as in panel b.
(e) A schematic representation of (d) is shown and has one nucleotide displacement to account for the difference of migration rate between the S1 nuclease-generated fragments and sequencing ladder (See Materials and Methods). The arrow indicates the proposed 3' end of YP1 gene.

points in a wide variety of genes (13,14).

A conserved sequence (5'-GG$^{C}_{T}$CAATCT-3'), often termed the CAT box, has been found 70-80 nucleotides upstream from the mRNA start points in a number of genes from higher eukaryotes (16). This CAT box was not found at the expected site in YP1 gene.

The 3' end of the YP1 mRNA was localized using the same procedure with the 3' end-labeled fragment B (Fig. 3a). The S1 resistant products of RNA protected fragment B are shown in Figure 3d. In Figure 3e the ends of the S1-resistant fragments are diagrammed alongside the nucleotide sequence of both complementary strands. This experiment indicates the approximate location of the 3' end of the YP1 messenger RNA is the cytosine marked by an arrow in this figure (position 1635 in Fig. 2). Twenty-six nucleotides upstream from this 3' end there is a sequence (5'-AATAAA-3') which has also been found 14-30 nucleotides upstream from the poly(A)$^{+}$ tract of most of the polyadenylated messenger RNAs that have been examined to date. Recently this sequence has been identified as part of the recognition site for polyadenylation in the late messenger RNAs of SV40 (17).

Location of the intron

The transcript map reported earlier (5) indicated that the YP1 gene has a single intron with boundaries (see Fig. 4a) approximately 100 nucleotides downstream (position 282 in Fig. 2) from a BamH I site and 165 nucleotides upstream (position 349 in Fig. 2) from the only BstE II site. From the DNA sequence shown in Figure 2 and the consensus sequence for intron splicing sites, 5'-AGGTPuAGT...PyNPyPyPyNPyAG-3' (where N is any nucleotide)(18,19), we can identify the likely boundaries for the intron. The sequence 5'-CTGTGAGT-3' (position 277 to 284 in Fig. 2) at the expected 5' boundary of the intron matches the consensus sequence at 6 out of 8 nucleotides and the

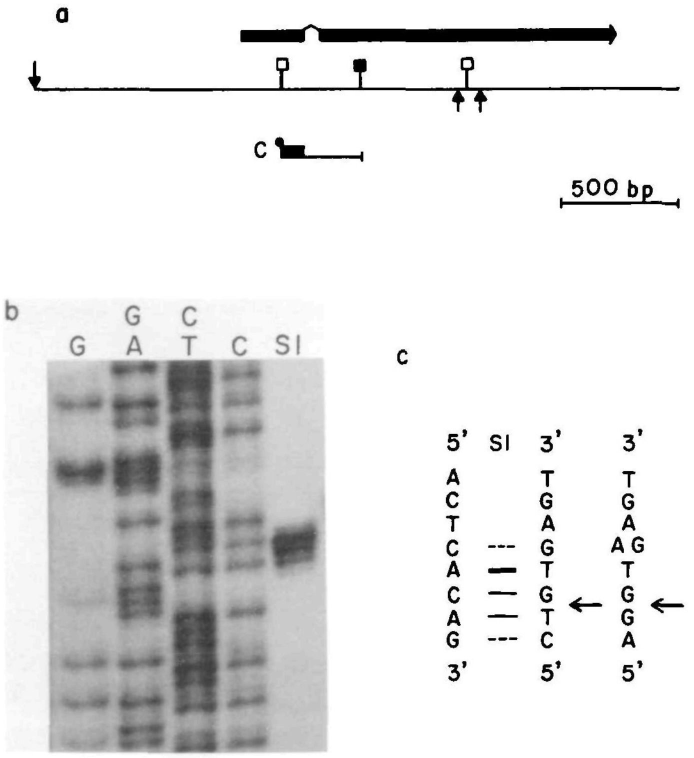

Figure 4: Location of the 5' end of the intron.
(a) A restriction map is shown. It uses the same symbols as in Figure 3 (a). Fragment C was used to map the 5' boundary of the intron.
(b) The S1-resistant material from fragment C was electrophoresed on an 8% polyacrylamide DNA sequencing gel. The amount of S1 nuclease in 200 µl reaction buffer was 50 units. In parallel is the DNA sequencing ladder prepared from the same fragment.
(c) A schematic representation of (b) is shown. The four columns from left to right are the sequence read from (b), the S1-resistant DNA fragments, the sequence of the non-transcribed strand and the consensus sequence of splicing, respectively. The sequencing standards are displaced by one nucleotide to account for their migration rate relative to the S1 nuclease-generated fragments (see text). The arrows indicate the splicing point.

sequence 5'-CCTTTGTAG-3' (positions 346 to 354 in Fig. 2) at the expected 3' boundary of the intron matches the consensus sequence at all 9 nucleotides. Since there are two other sequences (269-276 and 273-280) in area of the expected 5' splice site that match the consensus splice site in 5 of 8 nucleotides, the location of the 5' boundary was determined relative to the sequencing ladder. Fragment C (see Fig. 4a) was tested and the S1-resistant

DNA was analyzed on an 8% polyacrylamide sequencing gel in parallel to the DNA sequencing ladder prepared from the same fragment (see Fig. 4b). Figure 4c shows the ends of the S1 treated fragments aligned with the nucleotide sequence and indicates that the sequence between 277 and 284 is the splice location. The most likely splicing point for the 5' end of the intron is marked by an arrow in this figure. Within the limits (position 339 to 359 in Fig. 2) of our localization of the 3' end of the intron (5), the sequence at 346 to 354 (see Fig. 2) is the only reasonable candidate for a match with the 3' consensus splice site (18,19). We conclude that this is the most likely 3' boundary. Note that of the three possible matches to the 5' boundary consensus sequence that were discussed above, only the one at 277 to 284 will keep the translation frame in phase. We conclude that the most likely intron end points are as shown in Figure 2. These splice points indicate that the intron is 76 nucleotides long.

The location of the ends of the mRNA and the intron predict a full length YP1 mRNA of 1559 nucleotides plus the length of the poly(A) tail. This length is in accord with our previous estimates of the YP1 mRNA length (5).

Predicted protein sequence

We have used the DNA sequence information to predict the sequence of the YP1 protein. Kozak has proposed (20) that eukaryotic ribosomes initiate protein synthesis at the AUG closest to the 5' end of an mRNA. The first AUG codon in the YP1 mRNA occurs 58 nucleotides from the 5' end of the message (+59 in Fig. 2). The DNA sequence in this region (5'-AACCATG-3') matches the consensus sequence (5'-$^{C}_{A}$APyCATG-3') for the site of translation initiation (13). Twenty-four nucleotides upstream from this site is a region (see Fig. 2) with partial complementarity to the 3' end of the eukaryotic 18S ribosomal RNA. Hagenbuchle *et al*. (21) have observed similar regions in other eukaryotic mRNAs and have speculated that, as in prokaryot, such regions may contribute to ribosome recognition of mRNAs. If the intron sequence is deleted, there is an open reading frame for translation that is 1317 nucleotides long and begins at the above mentioned ATG (+59) and ends at TAA (+1452). The YP1 gene, therefore, can code for a protein 439 amino acids long (see Fig. 5) and approximately 48,300 daltons in molecular weight. This molecular weight estimate is close to the molecular weight of 46,000, estimated by SDS gel electrophoresis of the protein (22).

Yolk proteins are secreted from the fat body into Drosophila's circulatory system. There is evidence (22,3) that the primary yolk protein translation product contains an N-terminal hydrophobic signal sequence typical sn-Pro-Met-Arg-Val-Leu-Ser-Leu-Leu-Ala-Cys-Leu-Ala-Val-Ala-Ala-Leu-Ala-Lys-Pro-Asn-Gly-Ar

sp-Asn-Ser-Val-Asn-Gln-Ala-Leu-Lys-Pro-Ser-Gln-Trp-Leu-Ser-Gly-Ser-Gln-Leu-Glu-Ala-Ile-Pr

eu-Asp-Asp-Phe-Thr-Ile-Glu-Arg-Leu-Glu-Asn-Met-Asn-Leu-Glu-Arg-Gly-Ala-Glu-Leu-Leu-Gln-Gl

yr-His-Leu-Ser-Gln-Ile-His-His-Asn-Val-Glu-Pro-Asn-Tyr-Val-Pro-Ser-Gly-Ile-Gln-Val-Tyr-Va

ys-Pro-Asn-Gly-Asp-Lys-Thr-Val-Ala-Pro-Leu-Asn-Glu-Met-Ile-Gln-Arg-Leu-Lys-Gln-Lys-Gln-As

ly-Glu-Asp-Glu-Val-Thr-Ile-Ile-Val-Thr-Gly-Leu-Pro-Gln-Thr-Ser-Glu-Thr-Val-Lys-Lys-Ala-Th

ys-Leu-Val-Gln-Ala-Tyr-Met-Gln-Arg-Tyr-Asn-Leu-Gln-Gln-Gln-Arg-Gln-His-Gly-Lys-Asn-Gly-As

sp-Tyr-Gln-Asp-Gln-Ser-Asn-Glu-Gln-Arg-Lys-Asn-Gln-Arg-Thr-Ser-Ser-Glu-Glu-Asp-Tyr-Ser-Gl

al-Lys-Asn-Ala-Lys-Thr-Gln-Ser-Gly-Asp-Ile-Ile-Val-Ile-Asp-Leu-Gly-Ser-Lys-Leu-Asn-Thr-Ty

rg-Tyr-Ala-Met-Leu-Asp-Ile-Glu-Lys-Thr-Gly-Ala-Lys-Ile-Gly-Lys-Trp-Ile-Val-Gln-Met-Val-As

eu-Asp-Met-Pro-Phe-Asp-Thr-Ile-His-Leu-Ile-Gly-Gln-Asn-Val-Gly-Ala-His-Val-Ala-Gly-Ala-Al

ln-Glu-Phe-Thr-Arg-Leu-Thr-Gly-His-Lys-Leu-Arg-Arg-Val-Thr-Gly-Leu-Asp-Pro-Ser-Lys-Ile-Va

ys-Ser-Lys-Asn-Thr-Leu-Thr-Gly-Leu-Ala-Arg-Gly-Asp-Ala-Glu-Phe-Val-Asp-Ala-Ile-His-Thr-Se

yr-Gly-Met-Gly-Thr-Pro-Ile-Arg-Ser-Gly-Asp-Val-Asp-Phe-Tyr-Pro-Asn-Gly-Pro-Ala-Ala-Gly-Va

ly-Ala-Ser-Asn-Val-Val-Glu-Ala-Ala-Met-Arg-Ala-Thr-Arg-Tyr-Phe-Ala-Glu-Ser-Val-Arg-Pro-Gl

lu-Arg-Ser-Phe-Pro-Ala-Val-Pro-Ala-Asn-Ser-Leu-Gln-Gln-Tyr-Lys-Gln-Asn-Asp-Gly-Phe-Gly-Ly

la-Tyr-Met-Gly-Ile-Asp-Thr-Ala-His-Asp-Leu-Glu-Gly-Asp-Tyr-Ile-Leu-Gln-Val-Asn-Pro-Lys-Se

he-Gly-Arg-Asn-Ala-Pro-Ala-Gln-Lys-Gln-Ser-Ser-Tyr-His-Gly-Val-His-Gln-Ala-Trp-Asn-Thr-As

sp-Ser-Lys-Asp-Tyr-Gln

Figure 5: edicted YP1 protein sequence. This is predicted from the DNA sequence of Figure 2.

of secreted proteins (23). From the predicted protein sequence shown in Figure 5, it is clear that most residues of the 20 N-terminal amino acids contain hydrophobic side chains.

Possible Secondary Structure

Dyad symmetry sequences are often associated with the protein:DNA interactions necessary for regulation of gene expression (24). For this reason we search for dyad symmetry sequences, in the regions close to the 5' and 3' ends of the gene. We used as criteria (see Material and Methods) that

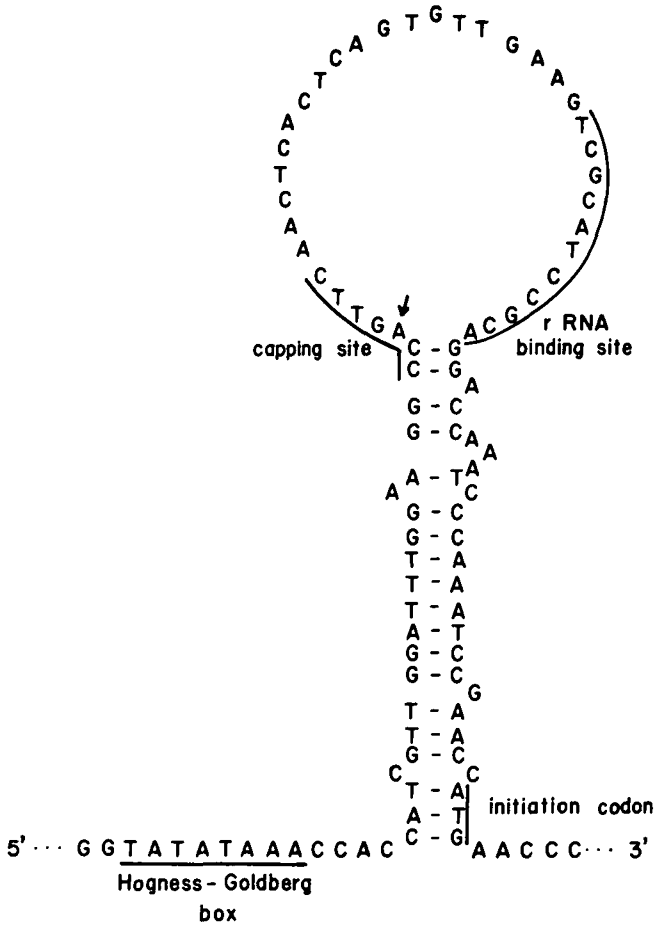

Figure 6: Likely secondary structure in the 5' region. The Hogness-Goldberg box, capping site, ribosome binding site and the initiation codon are shown. The 5' end of the mature transcript is marked by the arrow.

the symmetrical sequences must be at least 70% homologous, have a stability of ≤ -10kcal (Borer et al.(25) and Tinoco et al.(26)) and also be separated by no more than 50 nucleotides. In the regions searched (-347 to +70 and +1452 to +1775), only one such dyad symmetry was detected. It is of high stability (-21 kCal/mole) and its two parts (-21 to -1 and 36 to 61) would form the structure shown in Fig. 6. An interesting feature of this structure is that the four consensus sequences at the 5' end of the gene are located at the junctions between the duplex and single-stranded portions of the structure (see Fig. 6). This suggests that the occurence of this structure in either RNA or DNA may affect the function of one or another of the consensus sequences. Since the 5' end of the primary transcript is unknown several possibilities occur to us. If transcription initiates to the 5' side of the stem, then the stability of the structure predicts that it would occur in the primary transcript. It would be likely to have dramatic effect on the rate of translation initiation at the duplexed start codon and this effect could be eliminated by denaturation after a processing cleavage at the capping site. If transcription initiation is at the capping site, this structure may influence the rate of transcription initiation. Clearly, functional tests are required to examine these possibilities. One observation demonstrates that this structure and not its sequence is conserved between the different coordinately regulated yolk protein genes. The 5' end of the YP2 gene has an almost identical structure at the same location in the gene (unpublished results).

ACKNOWLEDGEMENTS

We thank Peter Gergen, Jim Haber, Howard Baum, Laura Kalfayan and Michael Garabedian for valuable discussions and for criticism of this manuscript. We also thank the Stanford MOLGEN project and the NIH SUMEX-AIM facility for allowing us to use the SEQ computer program, the National Institutes of Health for a research grant (P.C.W.) and the Gillette Company for a fellowship (M-C.H.).

Note Added in Proof: After this work was completed Hovemann et al. [Nucl. Acids Res. (1981) 9,4721-4734] published essentially the same sequence for the YP1 gene. The two groups worked independently and in ignorance of each other's work. Since we now know their result, we are able to compare the two results. There are only three differences. Two are in the intron: we find a stretch of 17A's at position 301-317 in our numbering system and they find

16A's; we find A at position 319 and they find a C. We sequenced both strands in this region. We have now reexamined the autoradiogram and find no ambiguity. The other differnce is at position 1634, where we find a T and they find a C. Our results for this region are shown in Figure 3, panels d and e.

REFERENCES

1. Warren, T.G. and Mahowald, A.P. (1979) Dev. Biol. 68, 130-139.
2. Handler, A.M. and Postlethwait, J.H. (1977) J. Exp. Zool. 202, 389-403.
3. Barnett. T., Pachl, C., Gergen, J.P. and Wensink, P.C. (1980) Cell 21, 729-738.
4. Barnett, T. and Wensink, P.C. (1981) UCLA-ICN Symposium, Developmental Biology using purified genes, in press.
5. Hung, M.C., Barnett, T., Woolford, C. and Wensink, P.C. (1981), submitted for publication.
6. Maxam, A.M. and Gilbert, W. (1980) In Method in Enzymology, 65, L. Grossman and K. Moldave, Eds. (New York: Academic Press) pp. 499-559.
7. Goodman, H.M. (1980) In Methods in Enzymology, 65, L. Grossman and K. Moldave, Eds. (New York: Academic Press), pp. 63-64.
8. Smith, H.O. and Birnstiel, M.L. (1976) Nucl. Acids Res. 3, 2387-2398.
9. Sollner-Webb, B. and Reeder, R.H. (1979) Cell 18, 485-499.
10. Queen, C. and Korn, L.J. (1979) In Methods in Enzymology, 65, L. Grossman and K. Moldave, Eds. (New York: Academic Press) pp. 595-609.
11. Ohmori, H., Tomizawa, J. and Maxam, A.M. (1978) Nucl. Acids Res. 5, 1479-1485.
12. Casey, J. and Davidson, N. (1977) Nucl. Acids Res. 4, 1539-1552.
13. Busslinger, M., Portmann, R., Irminger, J.C. and Birnstiel, M.L. (1980) Nucl. Acids Res. 8, 957-977.
14. Corden, J., Wasylyk, B. Buchwalder, A., Sassone-Corsi, P., Kedinger, C. and Chambon, P. (1980) Science 209, 1406-1411.
15. Hagenbuchle, O. and Schibler, U. (1981) Proc. Natl. Acad. Sci. USA 78, 2283-2286.
16. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucl. Acids Res. 8, 127-142.
17. Fitzgerald, M. and Shenk, T. (1981) Cell 24, 251-260.
18. Seif, I., Khoury, G. and Dhar, R. (1979) Nucl. Acids Res. 6, 3387-3398.
19. Sharp, P.A. (1981) Cell 23, 643-646.
20. Kozak, M. (1978) Cell 15, 1109-1123.
21. Hagenbuchle, O., Santer, M., Steitz, J.A. and Mans, R. (1978) Cell 13, 551-563.
22. Warren, T.G., Brennan, M.D. and Mahowald, A.P. (1979) Proc. Natl. Acad, Sci. USA 76, 2848-2852.
23. Blobel, G. and Dobberstein, B. (1975) J. Cell Biol. 67, 852-862.
24. Dickson, R.C., Abelson, J., Barnes, W.M. and Reznikoff, W.S. (1975) Science 187 27-34.
25. Borer, P.N., Dengler, B., Tinoco, I., Jr., Uhlenbeck, O.C. (1974) J. Mol. Biol. 86, 843-853.
26. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Unlenbeck, O.C., Crothers, D.M., Grallan, J. (1973) Nature New Biol. 246, 40-41.