Sequence homologies among the three yolk polypeptide (*Yp*) genes in *Drosophila melanogaster*

Y.L.Yan, C.J.Kunert and J.H.Postlethwait

Department of Biology, University of Oregon, Eugene, OR 97403, USA

Received September 9, 1986; Revised and Accepted December 2, 1986

ABSTRACT

To identify candidates for cis-acting sequences that regulate the sex-, stage-, and cell-specific expression of three coordinately regulated yolk polypeptide genes (*Yp*) in *Drosophila melanogaster*, we have mapped the *Yp3* transcript, sequenced a 4278 bp DNA fragment containing the *Yp3* gene, compared *Yp3* region sequences to corresponding parts of *Yp1* and *Yp2*, and compared the predicted amino acid sequence of YP3 to YP1 and YP2. The results showed that the *Yps* are largely homologous in translated regions, especially in the 3' half of the genes. Untranscribed flanking regions had little homology. A conserved inverted repeat (the H-box) has homology both to vertebrate steroid hormone receptor binding sites and to the ecdysone control region of *Drosophila's hsp23*. These results identify sequences to mutate in order to define elements that regulate *Yp* gene expression and govern YP polypeptide function.

INTRODUCTION

Sets of related genes whose members are expressed in the same tissues and developmental periods often contain sequence elements that reappear in each member of the set, and that may drive the coordinated pattern of expression (1-4). One set of developmentally regulated genes in *Drosophila melanogaster*, the three yolk polypeptide genes (*Yp1*, *Yp2*, and *Yp3*), encode the three major polypeptide precursors of egg yolk, or vitellogenin (5-7). All three *Yp* genes are X-linked: *Yp1* and *Yp2* are separated by 1225 bp and are divergently transcribed, while *Yp3* is about 1000 kb nearer to the centromere (8-10). Adult female flies make YPs in fat body cells and in ovarian follicle cells (11), beginning at about the time they eclose from the puparium; males normally do not make YPs (5-7,12).

At least two kinds of trans-acting factors influence *Yp* expression. First, hormones regulate *Yp* gene activity: the steroid hormone 20-hydroxyecdysone (20-OHE) causes *Yp* transcripts to accumulate in female fat body, and juvenile hormone stimulates development of ovarian follicles to the stage at which *Yp* genes are transcribed (13). High doses of 20-OHE, but not juvenile hormone, can induce *Yp* transcript accumulation in adult male fat body cells (12). We wanted

to learn whether Yp genes contain DNA sequences homologous to sequences found to bind to vertebrate steroid hormone receptors (14). The second kind of trans-acting influence on Yp expression is exerted by sex differentiation genes. For example, mutations in transformer genes can abolish Yp expression in diplo-X chromosomal females (12,15). We wonder if there exists a cis-acting regulatory element sensitive to the sex differentiation gene products -- a "sex box".

In this paper we have mapped the transcript of the Yp3 gene; sequenced a 4278 bp genomic fragment that includes the Yp3 gene; compared Yp3 sequences to Yp1 and Yp2 (16-19) to identify candidates for cis-acting regulatory elements involved in stage, cell, and sex specific expression; and compared the predicted sequence of the YP3 polypeptide to YP1 and YP2. We find a set of conserved DNA sequences that have features expected of cis-acting regulatory elements. These findings direct our attention to regions whose function in regulating cell, sex, and stage specific expression should be tested.

MATERIALS AND METHODS

Genotypes: We sequenced the Yp3 gene of Drosophila melanogaster, Canton S strain, isolated by Barnett et al. (9). Oregon R wild type flies provided RNA where indicated.

Sequence Determination: Restriction fragments of Yp3 were subcloned into M13 (20) and the DNA sequence was determined by the dideoxy method (21) using an M13 primer. For part of the region, we sequenced a set of nested deletions (22). New England Biolabs, Bethesda Research Laboratories, and Boehringer Mannheim Biochemicals supplied restriction and modification enzymes and deoxy and dideoxynucleotides. ³²P-nucleotides were from New England Nuclear.

Transcript Mapping: Total RNA from 2 to 3 day old adults was extracted by the Li/urea method and separated by electrophoresis in agarose-formaldehyde gels (23). Gels blotted to nitrocellulose were hybridized using 1x10⁷ cpm of probe per blot. Radioactively labelled DNA probes were produced by nick-translation (24). Single stranded radioactive probes were made from Yp3 sequences in M13 DNA as described (25).

Nuclease protection studies followed standard protocols (24). 50 µg of total RNA from adult female flies was hybridized to 1x10⁶ cpm ³²P-labelled single-stranded probe and the DNA/RNA duplex digested with S1 nuclease. DNA fragments protected by RNA from S1 digestion were electrophoresed on a standard 6% polyacrylamide-urea sequencing gel.

Sequence Analysis: DNA sequences were analyzed and compared using the molecular genetics programs of the University of Wisconsin Genetics Computer Group (UWGGC) running on a VAX computer.

RESULTS AND DISCUSSION

Mapping the Yp3 Transcript

The first steps toward determining the structure of the Yp3 gene were to characterize the size of the message and to identify the direction of transcription. The Yp3 transcript appeared on northern blots of adult female RNA as a single band about 1750 bp long (Fig. 1A). To identify the direction of

Figure 1. Size of the Yp3 transcript and the direction of transcription. (A) Transcript size. Lane 1: 10 µg of RNA from 2- to 3-day-old females electrophoretically separated and blotted onto nitrocellulose. Lane 2: size standard; 2 ng pBR322 DNA, one half of which had been digested with *Ava*I and *Xmn*I and the other half with *Acc*I and *Pvu*I. Fragment sizes in base pairs are indicated beside the lanes. The northern blot was probed with ³²P-labelled pYP3. (B) Transcript direction. Blot hybridization of RNA probed with either of two complementary single-stranded ³²P-labelled DNA fragments as shown in (C). Lanes 1 and 4: 1 ng of the BglII-Hind III fragment of Yp3. Lanes 2 and 5: 10 µg total RNA from adult males. Lanes 3 and 6: 10 µg of total RNA from females. Lanes 1 to 3 were probed with mp8YP3BH; lanes 4 to 6 were probed with mp9YP3BH. (C) Probes and conclusion. Yp3 is transcribed from BamHI towards HindIII.

Figure 2. S1 nuclease mapping of the *Yp3* transcript. (A) Fragments protected from S1 nuclease digestion by hybridization to *Yp3* transcripts. (B) Fragments used and protected. Indicated fragments of *Yp3* were subcloned into M13 and single-strand DNA probes corresponding to these genomic fragments were labelled with ^{32}P by running sequencing reactions in the absence of dideoxy nucleotides. The labeled fragments were hybridized to adult female total RNA and digested with the single-strand nuclease S1. DNA fragments protected from

the nuclease by hybridization to Yp3 transcript were electrophoresed in a sequencing gel. In (A), lanes G, A, T, and C show the sequence of the BglII-PstI fragment from mp8YP3 as a size standard. In lanes a to g, the fragments designated a to g as shown in (B) were used for hybridization. Protected fragments are indicated in (B) by thick dark lines, unprotected fragments by thin lines. Numbers indicate sizes in base pairs. Multiple bands sometimes occur due to "breathing" of the DNA/RNA duplex at the ends of the molecule and subsequent nibbling by the S1 nuclease.

transcription, two complementary single stranded probes were synthesized from opposite strands of the BamHI-HindIII fragment of Yp3 inserted into phage M13 mp8 or mp9 (called mp8Yp3BH and mp9Yp3BH) (Fig. 1C). These two probes were hybridized to electrophoretically separated adult male and female RNAs blotted onto nitrocellulose. The double stranded BglII-HindIII fragment of pYP3 served as a hybridization control (Fig. 1B lanes 1 and 4). The results showed that the mp9Yp3BH probe hybridized to the Yp3 transcript, but the mp8Yp3BH probe did not (Fig. 1B lanes 3 and 6). No transcript from males hybridized to either probe (Fig. 1B lanes 2 and 5). This result shows that Yp3 transcribes in the direction from BamHI towards HindIII (opposite to that previously reported (26)).

S1 nuclease protection studies localized the introns and exons of Yp3. The seven Yp3 fragments shown in Fig. 2B were cloned into M13 and used to make radiolabeled single stranded DNAs. Yp3 mRNA hybridized to the ssDNAs and protected complementary regions from digestion by the single strand specific nuclease S1. Sizes of the protected DNA fragments indicated the boundaries of exons and introns. A sequencing ladder provided size standards. The results

Figure 3. Sequenced fragments. The top line represents the SalI/HindIII DNA fragment containing Yp3. The thick arrow is the transcribed region, with the 5' region to the left. Translated regions are shown cross-hatched. The arrows below indicate the fragments sequenced, with the arrowheads indicating their 3' end.

-1823 GTGACGAGGGGCTCGAATCAAGGTTGTGCGCCGAGGAGTTGCTCCGCTGGGCGCGGAGA -1762

-1761 GATCATCTTCAGGTGCGGTGCGAAGAGCCCTGCGTCCATACAAATTCAGTCAAGGGCATGTTAAACGATCCGG -1683

-1682 GGCATCTGTTATGCTCGAAAGTCTCTCGCGGATGCCATTCGACCCGAGGCGGCTAAGGGATGCATGCTAAGT -1604

-1603 TCTCGCGGATGTCTATATTATACAGATCAGAACAAAGCAGATGTCGGGTAAATTCGCGGATTTGGCATCTGTCT -1525

-1524 GATTGCGGAGACTACGGACGCTGTGTGCTTGGCGCGGATTCGCTGTTCCAAACAAAGGAGGATGCGTGAATGCTG -1446

-1445 AAAGCATTATGGAGTAGTCTGTTTTTATATTATCTTGTAAATTCCTTACAGGAGATACACCATCATACCGAGA -1367

-1366 ATCTGGGCAAGAGGTGGCATGCGTTCCTAGACGAGATTTATGCGGTGCTGTGAGCTCCAGCTATCAATGCGT -1288

-1287 GCGCGCACTCTATATAGCATGCGGCGAAAGCATGTCTCCAAATTCCTACCGGTATGTCTGAGGATGTAATGAA -1209

-1208 CATGTTGATTATATGTTGATTACAGTGGCTGTGCAATACACCGTTCACTTCTGCAACATTGCGCGACTCTT -1130

-1129 CTGTCACCTTCAGCTGAGGATCCGAGGCGGAGACGAGACGAAACGAGGATGCGTGTGCCAAGAGGT -1051

-1050 CTTATGGCATGTGTTGGAAATGCGTACAAATATAAATAGCGCGTTATATAACCGTTCTGAGCTAAACTGTGA -972

-971 AGAATAACATTATATATTTTTTGTATACATATTAATATTAATTTACTATGTAACTATAGTAAATCAAT -893

-892 GTAGATAAGGTGTGTGATTAAAGCCTAAATATTATGTTACCTTTTTTTTACCATCAATTCCTCTTTTAT -814

-813 GCATGCAATTTTATGCACTTAAGCAGGACATTATTGTTATGATGAGACAAACATAATATATATATCTTTT -735

-734 TGGTGATGTGCTATGTTTGTGTTGAGCTCATCTTTAGCGTGTGCTATGCTTTTGCATATATAATATATGCAAT -656

-655 CACCTGGCGGCTGCTGATTGATCCAAATTTGGCGGCTTCGAATGCGTCAATGCGCGTCAACAGTTCTC -577

-576 ACTTGACGAGGTGTGCAAGTTTGTGCGAGTTCAATTCATCAAGGGATCTGCAAGTTGTTCAATCAATCGT -498

-497 ACTAGATACATTTTAAATGACAGACAAAATTTGCAATCTTTGGGATATATGCAATAATCTGAAGTGTGTT -419

-418 AAAACCAATGATGATGATCAAAATATATCAGTGCATACTACTAGTGTGCAAGTTTTCCTAAATCTAAAT -340

-339 CTTACATTTTATGATGATTTTTTTTGCACATTTTTTGCAGTGTGAGGCTCAAAACCTAGTGCATAG -261

-260 AGAATTAATGTTGGCAACACACACATGTGAATTAATCCGCTATTTGCAATCAATTTCCCTGACTGCACT -182

-181 TTATACCGCGGACGATCAGCAGAACGAAAGGGTGGGAAAAAATCGAGGCTAGACGCGGACGACGACAA -103

-102 CGACGACGACGACGACCTTCTGTGTGTCAGCAGAAATGCTGGCGTGCCTATCGGAGTCCGAGCTATTAAGC -24

-23 CAGAGTGGGCTGAAGGAGCCATCAACGCTGTTAGCGTTTGGCTGATCTGATTCATTCGAGATTGACCAAA 56

57 ATG AGT CTA AGG ATT TGC CTG CTG GCC ACC TGC CTC CTG GTG GCG GCC CAT GCC TCC 116
Met Met Ser Leu Arg Ile Cys Leu Leu Ala Thr Cys Leu Leu Val Ala Ala His Ala Ser

117 AAG GAT GCC TCC AAT GAC CGA CTG AAG CCG ACC AAG TGG CTG ACC GCC ACC GAG CTG GAG 176
Lys Asp Ala Ser Asn Asp Arg Leu Lys Pro Thr Lys Trp Leu Thr Ala Thr Glu Leu Glu

177 AAC GTG CCC TCC CTC AAC GAC ATC ACC TGG GAG CGT TTG GAG AAT CAG CCG CTG GAG CAG 236
Asn Val Pro Ser Leu Asn Asp Ile Thr Trp Glu Arg Leu Glu Asn Gln Pro Leu Glu Gln

237 GGC GGC AAG GTG ATC GAG AAG ATC T GTGATGAAACCGATGTTGCTGGAAATCTCCAGATTAACCTCC 306
Gly Ala Lys Val Ile Glu Lys Ile

307 TTGTGATCAGACTAG AC CAC GTT GGC CAA ATC AAG CAC GAT CTG ACC CCC AGC TTT GTG 367
Tyr His Val Gly Gln Ile Lys His Asp Leu Thr Pro Ser Phe Val

368 CCC AGC CCG AGC AAT GTG CCC GTC TGG ATT ATC AAG TCC AAT GGA CAG AAG GTT GAG TGC 427
Pro Ser Pro Ser Asn Val Pro Val Trp Ile Ile Lys Ser Asn Gly Gln Lys Val Glu Cys

428 AAG TTG AAC AAC TAT GTG GAG ACG GCC AAG GCA CAG CCC GGA TTC GGC GAG GAT GAG GTC 487
Lys Leu Asn Asn Tyr Val Glu Thr Ala Lys Ala Gln Pro Gly Phe Gly Glu Asp Glu Val

488 ACC ATT GTC CTG ACT GGT CTG CCC AAG ACC GGC GCT CAG CAG AAG GCC ATG GCG AGC 547
Thr Ile Val Leu Thr Gly Leu Pro Lys Thr Ser Pro Ala Gln Lys Ala Met Arg Arg

548 TTG ATC CAG GCC TAC GTC CAG AAG TAC AAC CTC CAG CAG CTG CAG AAG ACC GCC CAG GAG 607
Leu Ile Gln Ala Tyr Val Gln Lys Tyr Asn Leu Gln Gln Leu Gln Lys Asn Ala Gln Glu

608 CAG CAG CAG CAG CTC AAG AGC AGC GAC TAC GAC TAC ACC AGC AGC GAG GAG GCC GCT GAC 667
Gln Gln Gln Gln Leu Lys Ser Ser Asp Tyr Asp Tyr Thr Ser Ser Glu Glu Ala Ala Asp

668 CAA TGG AAA TCC GCC AAG GCT GCC AGC GGC GAT TTG ATC GTAATTGGTCGATTCTTATATTTCA 733
Gln Trp Lys Ser Ala Lys Ala Ala Ser Gly Asp Leu Ile

734 TAATTAAAGCTGACATATGATATTATTAATCAATTCAGATTCAG ATC ATT GAC CTC GGC TCC ACC CTG 802
Ile Ile Asp Leu Gly Ser Thr Leu

803 ACC AAC TTC AAA CGC TAC GCG ATG CTG GAT GTT CTG AAC ACC GGC GCC ATG ATC GGC CAG 862
Thr Asn Phe Lys Arg Tyr Ala Met Leu Asp Val Leu Asn Thr Gly Ala Met Ile Gly Gln

863 ACC CTG ATC GAT CTG ACC AAC AAG GGT GTG CCC CAG GAG ATC ATC CAT CTG ATC GGC CAG 922
Thr Leu Ile Asp Leu Thr Asn Lys Gly Val Pro Gln Glu Ile Ile His Leu Ile Gly Gln

923 GGA ATC AGC GCC CAT GTG GCC GGA GCT GCT GGC AAC AAG TAC ACC GCC CAA ACC GGA CAC 982
Gly Ile Ser Ala His Val Ala Gly Ala Ala Gly Asn Lys Tyr Thr Ala Gln Thr Gly His

983 AAG CTG CGC CGC ATC ACC GGT CTG GAT CCC AAG GTG CTG TCC AAG CBT CCC CAG ATC 1042
Lys Leu Arg Arg Ile Thr Gly Leu Asp Pro Ala Lys Val Leu Ser Lys Arg Pro Gln Ile

1043 CTG GGT GGT CTG TCC CGC GGC GAT GCT GAC TTC GTT GAT GGC ATT CAC ACA TGG ACC TTC 1102
Leu Gly Gly Leu Ser Arg Gly Asp Ala Asp Phe Val Asp Ala Ile His Thr Ser Thr Phe

1103 GCC ATC GGC ACG CCC ATC CBT TGC GGC GAT GTT GAC TTC TAC CCC AAC GGA CGS TCC ACC 1162
Ala Met Gly Thr Pro Ile Arg Cys Gly Asp Val Asp Phe Tyr Pro Asn Gly Pro Ser Thr

1163 GGT GTT CCC GGC TCC GAG AAT GTG ATC GAG GCT GTG GCC CBT GCC ACC CBT TAC TTT GCC 1222
Gly Val Pro Gly Ser Glu Asn Val Ile Glu Ala Val Ala Arg Ala Thr Arg Tyr Phe Ala

1223 GAG TCT GTG CBT CCC GGT AGC GAG GGC AAT TTC CCC GGT GTT CCG GCC AAC TCG CTG AAG 1282
Lys Ser Val Arg Pro Gly Ser Glu Arg Asn Phe Pro Ala Val Pro Ala Asn Ser Leu Lys

1283 CAG TAC AAG GAG CAG GAT GGC TTT GGC AAG GGC GGC TAC ATG GGT CTC CAG ATC GAC TAC 1342
Gln Tyr Lys Glu Gln Asp Gly Phe Gly Lys Arg Ala Tyr Met Gly Leu Gln Ile Asp Tyr

1343 GAT CTG CGC GGT GAC TAC ATC TTG GAG GTC AAC GGC AAG AGC CCC TTC GGT CAG CGC AGC 1402
Asp Leu Arg Gly Asp Tyr Ile Leu Glu Val Asn Ala Lys Ser Pro Phe Gly Gln Arg Ser

1403 CCT GGC CAC AAG CAG GCC GGC TAC CAT GGC ATG CAC CAC GGC CAG AAG TAG AGCGCCATGG 1464
Pro Ala His Lys Gln Ala Ala Tyr His Gly Met His His Ala Gln Asn End

1465 CCACGCCCCCTGGTTACCGGACCTTGGATGCTCACGCACTTTCTGATATCAGAAAAAACCAGGATGCTAG 1543

1544 TTATGCTAGAGTTTCATCAACCAATCAAAAAAGAAAAATCTATAAAATCCCATAAAAATAGCTGCATTTTCC 1622

^

1623 AAAAACTGAGTTTTTATAGCAATAGCATGGTTATTTCTGGATTGGATTCTATCTTTTATGGTATTAACACACA 1701

1702 CAGAAATTTGCTGGGCACATTTTATAGCCACCCCTTCTGAGTAATAGAAAAATTTCCGAAATATACATATTACAT 1780

1781 AGTAATTCGCCCCAACCTTAATGAGCTATATATAAAAGATTAATGATATATACAGGTGATCTTAGCAGATGC 1859

1860 TTATCCAAAAACACACGATTAATTAAGCAACAAAAAGTGGCTAAAAATACATTTATGACCACTAATGAAGGTACAC 1938

1939 GAAAGAAATGTAGATATTAATTAACCTGAAGAAATTAGGAATTAATCTATATAAAATCAAAATTTAGAAAATGTGCA 2017

2018 GCTTGTATTACTAGCAACCTAGATGCTTACACAGATTGCGAAGTTGGGATGGAAATACGACACAGAGTGGATCA 2096

2097 TTGATGGGCGAAGTGAGAGTGAGGCACTAGTGTGGTGGCACTTGATGTGCACTCAATTAACCTTGCAATCGGT 2175

2176 TTATCGTTAGTGACTACTGTTCAAAAATCACTGGGCAACCTGTGTAACTCAATTTGCTTACAGTTTGGGACATG 2254

2255 CGCGGTGTAATGTCAAGTTGACCTTTATCAATGCAATAGACAACTAGAAAGGCGACGAAACAGCAGAGTGGAA 2333

2334 AATAGCGGAGATAGGGAGCTGGAGTGACAGGACGGAATGACACAGTTGGCGCTTTTGTGTCATGCTGTCACA 2412

2413 TGTTGCTTGACTCTGACGAGCAAGATGGCGCTTAAGCTT 2455

Figure 4. Nucleotide sequence of the 4278 bp region containing *Yp3*. The predicted amino acid sequence is indicated below the DNA sequence. Carets at +1 and +1622 mark the beginning and end of the transcript as determined in Fig. 2. The TATA box homology and AATAAAs are indicated by dashed underlines. Predicted translation begins at the first AUG in the mRNA and continues to the first in frame stop codon. Intron positions are as determined from Fig. 2.

(Fig. 2A) showed that Yp3 contains three exons of 261, 383, and 844 base pairs, and two introns of 62 and 72 base pairs (Fig. 2B). The first intron of Yp3 is in a homologous position to the single intron of Yp1 and Yp2, but the second intron of Yp3 is not found in Yp1 or Yp2 (26).

The Sequence of the Yp3 Gene

Figure 3 shows the sequenced DNA fragments. Figure 4 shows the sequence of 4278 bp of genomic DNA containing Yp3 and flanking regions. The probable cap site of Yp3 according to the S1 mapping data is indicated as +1. The 5' terminus of the Yp3 gene has the sequence A T C A A A C, a 5/7 match to the insect cap site consensus sequence A T C A G/T T C/T (27). A 5' untranslated leader of 56 base pairs precedes the first ATG triplet. Since this ATG is embedded in a context (C C A A A T G A) similar (6/9) to that found in other eukaryotic genes (C C A/G C C A T G G) (28,29), we assume in the absence of any protein sequencing data that this is the initiation codon. The 5' and 3' splice sites of both introns determined from the S1 mapping data also conform to consensus sequences (↓ G T A/G A G T-----Y₁₁ N C/T A G ↓) (30,31). In addition to the splice site consensus sequences, both introns of Yp3 contain renditions (4 matches in 5 bp) of the internal splice signal C/T T A/G A C/T that usually appears 18 to 35 bp upstream of the 3' splice site (32) (intron 1: A T A A C - 20 bp - T A G ↓; intron 2: T T T A T - 16 bp - C A G ↓).

An in-frame TAG stop codon appears at +1451. The 168 bp stretch between this stop codon and the 3' end of the Yp transcript contains three other stop codons in phase, and five stop codons out of phase. The G at +1622 is the polyadenylation site according to the S1 mapping data. This G occurs in the context G↓A A A A, which fits the poly(A) addition site consensus sequence N↓A A A (33). Twenty base pairs preceding the polyadenylation site is a copy of the consensus polyadenylation signal A A T A A A. In most genes encoding polyadenylated messages, this signal precedes the poly-A site by 11-30 bp (34) and is essential for cleavage (35,36). While this A A T A A A sequence is the second of five within about 450 bp at the end of the Yp3 gene, it is the only one that is used according to the S1 transcript mapping data.

Conserved sequences downstream of the A A T A A A may be required for correct processing of the 3' end of eukaryotic messages (37-39). While the consensus sequences C A C/T T G (33) or C/T G T G T T C/T C/T (40) found frequently within a few bases of the poly(A) addition site are not present in the 3' region of Yp3, a G/T cluster found in many vertebrate genes (37) appears just eleven nucleotides downstream of the poly(A) addition site of Yp3 (Table 1). This sequence A A G T T T T T has 8/9 matches to the sequence A G G T T T T T

Table 1. G/T clusters following poly(A) addition sites of *Yp* genes.

Status	Location	Signal	Sequence	Reference
Yp1				
a (used)	+1603	AATAAA	CGCTCGATAAAGAGCTTAAAAATATcTAT <u>TTTGT</u> TTATCTTAAATCATTTAGGAACATAATAGTCTAATAGATCATCCCAAAAAA	(16)
Yp2				
a (used)	+1589	AATAAA	AAACGTTTGCAATTTTAAACATTCTAAAGATTCAGTTCATATCGGAAAAAACCCAGTTCATTTACAATAAAAAACATTTGCTT	(18)
b (used)	+1665	AATAAA	AACAATTGCTTATGTCGAATATTTGAGATTCCTCAATGCTCCTATATAAAAAATATCCAAACCAATATGCAATGCCCATGAGG	(18)
Yp3				
a (unused)	+1522	AATAAA	AACCCGGAATGCGTATTTAGCTTAGAAGTTTCATCAACAATCAAAAAAGAAAAATCTATAAAATCCCATAAAAATAAAGCTGCA	
b (used)	+1602	AATAAA	AGCTGCAAAATTTTCGAAAAGTCAAGTTTAAATAGCAATAGCAATGGTTATTCTGGATTGGATTCTAACTTTTATGGTATTAAAAA	
c (unused)	+1814	AATAAA	AGATATAATGCATATATCACAGGTGATCTTAAGCAGATGCTTAACCAAAAAACAACAGATAAATAAGCAACAAAAAGTGCCTAA	
d (unused)	+1882	AATAAA	GCAAAACAAAAAGTGCCTAAATACAATTATGACACCTAATGAAAGTACACGAAAGAAATGTAGATAATAATAAATACTGAAAAGAA	
e (unused)	+1955	AATAAA	TAAACTGAAAAAGAAATTAGGAATAACTCATAAAAATCAAAATTTAGAAAACTGTGAGCTTGGTATTACTAGCACCCCTAGATGCTT	

The small letters and arrows indicate the poly(A) addition sites according to S1 mapping data. G/T clusters (37) are underlined. The *Yp3* data are from this paper.

shown to be an important signal in the location of 3' ends in SV40 (41). *Yp1* and *Yp2* also have a G/T cluster within eleven bp of the poly(A) addition site (Table 1). All of the four poly(A) addition sites of *Yp1,2* and *3* have a G/T cluster beginning within 31 bp of the A A T A A hexanucleotide. Of the five A A T A A sequences in *Yp3*, only two have a G/T cluster within 31 bp of the A A T A A, and only one of these two is used. Evidently the presence of a G/T cluster downstream of an A A T A A motif is insufficient to specify the site of 3' cleavage and poly(A) addition for the *Yp3* transcript. Further in vitro mutagenesis experiments and transformation experiments are needed to determine sequence elements that specify the exact poly(A) addition site.

As in other genes, the exons of *Yp3* tend to be G+C rich, while the introns and flanking regions tend to be A+T rich (3,42). A 15 bp long stretch of A+T occurs at position -675. While the significance of such A+T rich regions is currently unknown, the receptors of some vertebrate steroid hormones bind preferentially to A+T rich DNA, and an A+T rich sequence has been implicated in the steroid-sensitive expression of small heat shock genes (43).

Sequence Homologies in 5' Flanking Regions

Matrix analyses (44) comparing *Yp1*, *Yp2* and *Yp3* revealed extensive homology in translated regions but little homology in transcribed but non-translated regions or in flanking DNA (Fig. 5). Since the 5' flanking DNA is likely to be involved in the regulation of *Yp3*, we analyzed the matrix plots for these regions in detail (Fig. 5).

Nine repeats of the triplet 5'-C G A-3' are clustered from -85 to -117 of *Yp3* (Table 2). This triplet is also concentrated in the corresponding region of *Yp1* and *Yp2* (16-19).

The heptamer T/A A T G C A A repeats three times tandemly at about -800 in *Yp3*, and this sequence or its complement appears four other times in the 5' flanking DNA of *Yp3* and five times in the region between *Yp1* and *Yp2* (Table 2).

Figure 5. Graphic matrix analysis (44) of yolk polypeptides. (A), (C), and (E): 1225 bp of 5' flanking DNA. Each point represents a position in which 14 of 21 base pairs are the same in the two compared genes. (B), (D), and (F): transcribed region. Each point represents a position in which 8 of 10 base pairs are the same in the two compared genes. (A) and (B): Yp1 and Yp3 compared. (C) and (D): Yp2 and Yp3 compared. (E) and (F): Yp1 and Yp2 compared. Long stretches of homology show up as diagonal lines.

Table 3. Homologies of Yp H-box to ECR of hsp23

hsp23 ECR	-200	5'-ATGGCAGAT-3' (ref: 50)
		: : : : : :
Yp H box consensus:		5'-ATCAGCAGAACTAAA-3'
hsp23 ECR:	-200	5'-ATGGCAGAT3'
		: : : : : :
	-254	5'-ATCAGGCGGCAGCAG\
Yp1 H box:		: : : : : : : : : : : :
	-225	3'-TACTGAATCGTCGTG/
		: : : : : :
hsp23 ECR:	-220	3'-TAAAAGGTA-5'
hsp23 ECR:	-228	5'-ATTTTCCAT-3'
		: : : : : :
	-87	5'-CGACTTCTGTGG\
Yp3 H box:		: : : : : : : : : : : :
	-61	3'-GCTAAAAGACGAC/
		: : : : : :
hsp23 ECR:	-220	3'-TAAAAGGTA-5'

Repeated sequence elements of a few base pairs occurring in other genes between -50 and -300 identify the binding sites of promoter-specific transcription factors (45), but their significance in *Yp* genes is as yet untested.

The matrix analysis (Fig. 5) identified repeats of close relatives of the sequence 5'-A T C A G C A G A A C T A A A-3' near all three *Yps*. This sequence seems likely to be important for *Yp* regulation on both structural and functional grounds. First, the element is conserved, repeated, and localized: it occurs three times between -50 and -300 bp in *Yp3*, twice between -150 and -190 in *Yp2*, and four times between -100 and -450 in *Yp1* (Table 2). Second, copies of this element (in *Yp1* at -254 and in *Yp3* at -75) are present adjacent to imperfect inverted complements (Table 3), a feature shared by other cis-acting regulatory elements (45,46). Third, a part of this sequence has homology (6/9) to binding sites of the glucocorticoid receptor in mouse mammary tumor virus (47) and other vertebrate steroid hormones (48); for convenience we call this the H-box. Fourth, the H-box has strong homologies (7/10) to the ecdysone response element of *Drosophila*'s hsp23 gene (49) (Table 3).

In addition to the structural evidence, functional data suggest that the H-box could be important in regulation of *Yps*. Mutations constructed near *Yp1* in vitro that delete H-boxes contained in either the 125 bp between -196 and -321 (50) or the 257 bp between -89 and -346 (51) stop expressing *Yp1* in fat body cells. Both deletions removed the H-box which had a nearby inverted complement, suggesting that future experiments should be designed to specifically disrupt this structure.

We compared the 125 bp fat body control sequence identified by Garabedian et

Figure 6. The predicted amino acid sequence of Yp3 compared to that of Yp1 (16,17) and Yp2 (18,19). The caret at 19 amino acids indicates the predicted site of signal peptide cleavage. Other carets indicate the sites of introns.

79

surrounding the H-boxes (underlined below) at about -250 to -300 of Yp1 and Yp3 and -500 of Yp3.

```

Yp3 -526  5'-ATCTGCACAAGT...TGTTTCAATCAATCCGTACTAGAATAC-3' -488
              ::::: :      ::::: ::::: : : ::::: :
Yp1 -313  5'-CGGTGCACAAC TACAATGTTGCAATCAG.CGGAAC TACAAAAGT-3' -272
              : : ::::: : : : : : : : : : : : : :
Yp3 -288  5'-AGAGGTTCAAAAACCTTAGTGCATAAG..AGAACTAAATGGT-3' -248

```

We predict that features of this sequence may be important controlling elements in Yp expression. Table 2 shows several other conserved regions brought to our attention by the matrix analysis, but their significance is unknown.

Predicted Protein Sequence

Removing the two introns from Yp3 gives an open reading frame 1260 nucleotides long beginning with an ATG methionine codon (+57) and ending with a TAG stop codon (+1454) (Fig.4). This sequence can encode a 420 amino acid long polypeptide (Figs. 4 and 6).

The first 19 amino acids have features expected of a signal peptide, a 15 to 30 amino acid tag that directs insertion of the polypeptide into the lumen of the endoplasmic reticulum (52). These features include: positively charged amino acids near the amino terminus (Arg-5); a hydrophobic domain of 10 to 15 amino acid residues (residues 6 to 17); one or two uncharged polar residues within the hydrophobic domain (Cys-7 and Cys-12); serine and threonine near the carboxy end of the hydrophobic region (Thr-11); and an alanine or glycine at the cleavage site (probably Ala-19) (53). von Heijne's (54) method for calculating the most likely signal sequence cleavage site predicts cleavage after Ala 19 (Fig. 6). A peptide of the appropriate size is in fact cleaved from the primary translation product of YP3 (17,55-57), however the exact amino-terminal amino acid of the mature form of YP3 has not yet been identified.

YP3, like YP1 and YP2, becomes processed by glycosylation, sulfation, and phosphorylation before it is secreted (56-58). Since the amino acid sequences of all three YPs lack the Asn-X-(Ser or Thr) site of N-linked glycosylation (59), glycosylation is likely to be O-linked on the YPs.

The tripeptide Arg-Gly-Asp appears at approximately the 300th amino acid residue in all three YPs, and a second time at 387 in YP3. This RDG tripeptide has recently been implicated in mediating the interaction of several proteins (such as fibronectin, vitronectin, and fibrinogen) with cell surfaces (60). Experiments now in progress are testing whether this sequence plays a role in recognition events at the oocyte cell surface that bring YPs into the oocyte.

As suggested by matrix analyses (Fig. 5), the predicted amino acid sequences

Table 4: Codon usage in Yp3.

amino acid	codon	%	amino acid	codon	%	amino acid	codon	%	amino acid	codon	%
Phe	TTT	30	Ser	TCT	4	Ala	GCT	17	Glu	GAA	0
	TTC	70		TCC	37		GCC	76		GAG	100
				TCA	0		GCA	2			
Leu	TTA	0		TCG	7		GCG	5	Cys	TGT	0
	TTG	14		AGT	4					TGC	100
	CTT	0		AGC	48	Tyr	TAT	7			
	CTC	17					TAC	93	Arg	CGT	33
	CTA	3	Pro	CCT	5					CGC	50
	CTG	67		CCC	73	His	CAT	36		CGA	6
				CCA	0		CAC	64		CGG	0
				CCG	23					AGA	0
Ile	ATT	20				Gln	CAA	10		AGG	11
	ATC	80					CAG	90			
	ATA	0	Thr	ACT	4				Gly	GGT	31
				ACC	83					GGC	50
Val	GTT	29		ACA	4	Asn	AAT	30		GGA	19
	GTC	21		ACG	8		AAC	70		GGG	0
	GTA	0									
	GTG	50				Lys	AAA	7			
							AAG	93	Stop	TAG	-
						Asp	GAT	55			
							GAC	45			

of YP1, YP2, and YP3 show substantial homologies (Fig. 6) (YP1/YP2 53%, YP1/YP3 53%, YP2/YP3 48%), reaching up to 73% in their carboxyl half (Fig. 6). This suggests that structural features necessary for common functions of the YPs, such as their transport in the blood, sequestration into the oocyte, packaging into yolk granules, and utilization during embryogenesis may be contained in this portion of the molecules. If the functions of the YPs overlap due to these structural homologies, it could explain why a mutation that blocks the production of mature YP2 is fully fertile (61).

Homologies in DNA and peptide sequences and in intron locations and numbers suggest an evolutionary relationship among the three Yps. Since intron loss occurs in evolution, but there is no proof of intron addition (62), we assume that the Yp1/2/3 precursor had two introns rather than one. The common Yp1/2/3 precursor with two introns then duplicated, with the Yp1/2 precursor residing at chromosome region 9A and the Yp3 precursor at 12A-D on the X chromosome. The Yp1/2 precursor lost an intron and then duplicated again, giving rise to Yp1 and Yp2 both at 9A. These events must have occurred early since all *Drosophila* species so far examined have two closely linked Yp's and a third Yp some distance away on the X chromosome (63).

Codon Usage in Yp3

Synonymous codons are not used with equal frequencies in Yp3 (Table 4). Thirty of the 61 amino acid coding triplets account for over 90% of the codons in Yp genes and six codons are not used at all, a bias seen in other *Drosophila* genes (64). Graphic analysis of codon preference (65) for Yp3 region DNA clearly defines the reading frames, introns, and exons of the gene (Fig. 7). The data in

Figure 7. Codon preference plot for Yp3 region DNA. The three reading frames for transcription toward the right are shown. The computer compared codons to a table of codon frequencies used by ten *Drosophila* genes and calculated an index of the likelihood a window of 25 codons would use those codons (65). AUG codons (up) and stop codons (down) are shown just below each plot. The hash marks represent the position of any codon used less than 6% of the time. The structure of Yp3 is shown just above the plot for the second reading frame. The plot defines the translated regions of the three exons of Yp3, in the third and first reading frames.

Table 4 show a substantial tendency for frequently-used codons to employ the more strongly hydrogen bonded G/C pairs, suggesting that stronger binding between a codon and its tRNA may optimize translation rate.

Infrequent codons are used disproportionately often in Yp signal peptides -- 14% of the 57 putative signal sequence codons but only 4.3% of the 1301 codons in Yp1,2 and 3 are used 6% or less of the times each codon's amino acid is specified. Rare codons are also frequent in peptide leader sequences in other proteins, and they may promote slow translation of the signal peptide, allowing time for the nascent polypeptide to interact with the secretory apparatus (66,67).

Conclusions

These studies direct our attention to DNA sequences that may be important in regulating the coordinated expression of Yp genes in a specific sex and in

specific tissues. A significant feature seems to be a repeated and conserved element (the H-box) with homologies to both vertebrate steroid hormone receptor binding sites and to the ecdysone control region of a small heat shock gene. The results further suggest nucleic acid signals that may specify the site of polyadenylation, and conserved amino acid sequences in the yolk polypeptides that may be important in directing their secretion from fat body and ovarian follicle cells, their transport in the blood, their recognition by the oocyte surface, their sequestration into the oocyte, and finally their utilization or function during embryogenesis. Experiments now in progress are testing the importance of these signals. In vitro mutagenesis to modify defined sequences, followed by introduction of mutant genes into the fly genome by P-element-mediated gene transfer allows a functional test of the sites suggested to be important by the structural studies reported in this paper.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Institutes of Health and the American Heart Association. The technical assistance of Denita Widders-Ellis and Juanita Postlethwait, and conversations with Don Kane are gratefully acknowledged. We thank Drs. M. Bownes and P. Wensink, who communicated their sequence of Yp3 to us after our manuscript was submitted for publication.

REFERENCES

1. Delaney, S.J., Smith, D.F., McClelland, A., Sunkel, C. and Glover, D. (1986) *J. Mol. Biol.* **189**, 1-11.
2. Spieth, J., Denison, K., Kirtland, S., Cane, J., and Blumenthal, T. (1985) *Nucl. Acids Res.* **13**, 5283-5295.
3. Wong, Y., Pustell, J., Spoerel, N. and Kafatos, F. (1985) *Chromosoma* **92**, 124-135.
4. Pelham, H. (1985) *Trends in Genetics* **1**, 31-35.
5. Kambyzellis, M.P. (1977) *Amer. Zool.* **17**, 535-549.
6. Bownes, M. and Hames, B.D. (1977) *J. Exp. Zool.* **200**, 149-156.
7. Postlethwait, J.H. and Kaschnitz, R. (1978) *FEBS Letters* **95**, 247-251.
8. Postlethwait, J.H. and Jowett, T. (1980) *Cell* **20**, 671-678.
9. Barnett, T., Pacht, C., Gergen, J.P. and Wensink, P.C. (1980) *Cell* **21**, 729-738.
10. Riddell, D.C., Higgins, M.J., McMillan, B.J. and White, B.N. (1981) *Nucl. Acids Res.* **9**, 1323-1338.
11. Brennan, M.D., Weiner, A.J., Goraliski, T.J. and Mahowald, A.P. (1982) *Develop. Biol.* **39**, 225-236.
12. Postlethwait, J.H., Bownes, M. and Jowett, T. (1980) *Develop. Biol.* **79**, 379-387.
13. Jowett, T. and Postlethwait, J.H. (1980) *Develop. Biol.* **80**, 225-234.
14. Shrader, W.T., Compton, J.G. and O'Malley, B. (1983) In Eriksson, H. and Gustafsson, J. (eds), *Steroid Hormone Receptors: Structure and Function*, Elsevier, Amsterdam, pp. 247-265.
15. Belote, J., Handler, A.M., Wolfner, M.F., Livak, K.J. and Baker, B.S. (1985) *Cell* **40**, 339-348.

16. Hung, M. and Wensink, P.C. (1981) *Nucl. Acids Res.* **9**, 6407-6419.
17. Hovemann, B., Galler, R., Walldorf, U., Kupper, H. and Bautz, D.K.F. (1981) *Nucl. Acids Res.* **9**, 4721-4734.
18. Hung, M. and Wensink, P.C. (1983) *J. Mol. Biol.* **164**, 481-489.
19. Hovemann, B. and Galler, R. (1982) *Nucl. Acids Res.* **10**, 2261-2275.
20. Messing, J. and Vieira, J. (1982) *Gene* **19**, 269-276.
21. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463-5468.
22. Dale, R.M.K., McClure, B.A. and Houchins, J.P. (1985) *Plasmid* **13**, 31-40.
23. Richards, G., Cassab, A., Bourouis, M., Jarry, B. and Dissous, C. (1983) *EMBO J.* **2**, 2137-2142.
24. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
25. Hu, N. and Messing, J. (1982) *Gene* **17**, 271-277.
26. Hung, M.-C., Barnett, T., Woolford, C. and Wensink, P. (1982) *J. Mol. Biol.* **154**, 581-602.
27. Hultmark, D., Klemenz, R. and Gehring, W.J. (1986) *Cell* **44**, 429-438.
28. Kozak, M. (1984) *Nucl. Acids Res.* **12**, 857-872.
29. Kozak, M. (1986) *Cell* **44**, 283-292.
30. Mount, S.M. (1982) *Nucl. Acids Res.* **10**, 459-472.
31. Teem, J.L., Abovich, N.A., Kaufer, N.F., Schwindinger, W.F., Warner, J.R., Levy, A., Woolford, J., Leer, R.J., van Raamsdonk-Duin, M.M.C., Mager, W.H., Planta, R.J., Schultz, L., Friesen, J.D., Fried, H. and Rosbash, M. (1984) *Nucl. Acids Res.* **12**, 8295-8312.
32. Keller, E. and Noon, W.A. (1985) *Nucl. Acids Res.* **13**, 4971-4981.
33. Berget, S.M. (1984) *Nature* **309**, 179-182.
34. Proudfoot, N.J. and Brownlee, G.G. (1976) *Nature* **263**, 211-214.
35. Wickens, M. and Stephenson, P. (1984) *Science* **226**, 1045-1051.
36. Swimmer, C. and Shenk, T. (1985) *Nucl. Acids Res.* **13**, 8053-8063.
37. Birnstiel, M.L., Busslinger, M. and Strub, K. (1985) *Cell* **41**, 349-359.
38. Kessler, M.M., Beckendorf, R.C., Westhafer, M.A. and Nordstrom, J.L. (1986) *Nucl. Acids Res.* **14**, 4939-4952.
39. Mason, P.J., Elkington, J.A., Malgorzata, M.L., Jones, M.B. and Williams, J.G. (1986) *Cell* **46**, 263-270.
40. McLauchlan, J., Gaffney, D., Whitton, J.L. and Clements, J.B. (1985) *Nucl. Acids Res.* **13**, 1347-1368.
41. Sadofsky, M., Connelly, S., Manley, J.L. and Alwine, J.C. (1985) *Mol. Cell Biol.* **5**, 2713-2719.
42. Vincent, A., Colot, H.V. and Rosbash, M. (1985) *J. Mol. Biol.* **186**, 149-166.
43. Hoffman, E. and Corces, V. (1986) *Mol. Cell Biol.* **6**, 663-673.
44. Maizel, J.V. and Lenk, R.P. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7665-7669.
45. Heberlein, U., England, B. and Tjian, R. (1985) *Cell* **41**, 965-977.
46. Ptashne, M. (1986) *A Genetic Switch*, Cell Press, Cambridge.
47. Payvar, F., DeFranco, D., Firestone, G.L., Edgar, B., Wrangé, O., Okret, S., Gustafsson, J.A. and Yamamoto, K.R. (1983) *Cell* **35**, 381-392.
48. Baskevitch, P.P. and Rochefort, H. (1985) In Sluysers, M. (ed), *Interaction of Steroid Hormone Receptors with DNA*, pp. 107-125, Ellis Horwood, Chichester.
49. Mestril, R., Schiller, P., Amin, J., Klapper, H., Ananthan, J. and Voellmy, R. (1986) *EMBO J.* **5**, 1667-1673.
50. Garabedian, M.J., Shepherd, B.M. and Wensink, P.C. (1986) *Cell* **45**, 859-867.
51. Postlethwait, J.H., and Kunert, C.J. (1986) In Ralph, C. (ed), *Comparative Endocrinology: Development and Direction*, Alan Liss, New York, pp. 33-52.
52. Walter, P., Gilmore, R. and Blobel, G. (1984) *Cell* **38**, 5-8.
53. Inouye, M., Inouye, S., Pollitt, S., Ghrayeb, J. and Lunn, C.A. (1985) In

-
- Current Communications in Molecular Biology, Gething, M.J. (ed) Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 54-61.
54. vonHeijne, G. (1983) Eur. J. Biochem. 133, 17-21.
 55. Warren, T.G., Brennan, M.D. and Mahowald, A.P. (1979) Proc. Natl. Acad. Sci. USA 76, 2848-2852.
 56. Minoo, P. and Postlethwait, J.H. (1985) Biochem. Genetics 23, 913-932.
 57. Brennan, M.D. and Mahowald, A.P. (1982) Insect Biochem. 12, 669-673.
 58. Baeuerle, P.A. and Huttner, W.B. (1985) J. Biol. Chem 260, 6434-6439.
 59. Lennarz, W.J. (1979) In Russell, T.R., Brew, K., Faber, H. and Schultz, J. (eds) From Gene to Protein: Information Transfer in Normal and Abnormal Cells. Academic Press, New York, pp. 407-423.
 60. Ruoslahti, E. and Pierschbacher, M.D. (1986) Cell 44, 517-518.
 61. Tamura, T., Kunert, C. and Postlethwait, J. (1985) Proc. Natl. Acad. Sci. USA 82, 7000-7004.
 62. Gilbert, S., Marchionni, M. and McKnight, G. (1986) Cell 46, 151-154.
 63. Dong, K., Craddock, E.M. and Kambyzellis, M.P. (1985) Genetics 110, s56.
 64. O'Connell, P. and Rosbash, M. (1984) Nucl. Acids Res. 12, 5495-5513.
 65. Gribskov, M., Devereux, J. and Burgess, R.R. (1984) Nucl. Acids Res. 12, 539-549.
 66. Spieth, J., Denison, K., Zucker, E. and Blumenthal, T. (1985) Nucl. Acids Res. 13, 7129-7138.
 67. Holm, L. (1986) Nucl. Acids Res. 14, 3075-3087.