Volume 12 Number 13 1984 Nucleic Acids Research

Sequence, structure, and codon preference of the *Drosophila* ribosomal protein 49 gene

Peter O'Connell[1] and Michael Rosbash[2]

[1]Howard Hughes Medical Institute, 732 Wintrobe Building, University of Utah Medical School, Salt Lake City, UT 84132, and [2]Department of Biology, Brandeis University, Waltham, MA 02254, USA

Received 15 March 1984; Revised and Accepted 8 June 1984

ABSTRACT

In this communication, we describe several features of the D. melanogaster gene which codes for ribosomal protein 49 (rp49). Nucleotide sequence analysis in conjunction with primer extension and S_1 nuclease protection experiments show that the structure of the rp49 gene consists of a 102 bp 5' exon, a single 59 bp intron, and a 420 bp 3' exon, encoding a total of 132 amino acids. The rp49 gene shares many features with other abundantly expressed Drosophila genes, including codon preference, which are discussed.

INTRODUCTION

The expression of eukaryotic ribosomal protein genes has been studied in a number of organisms and in a number of experimental situations (1-4). The evidence accumulated to date suggests that the synthesis of ribosomal proteins is coordinately controlled, i.e., the synthesis of most ribosomal proteins changes in concert with the physiological or developmental state of the cells or organism (e.g., 5-9).

Ribosomal protein genes or mRNAs have been cloned from a number of eukaryotes (e.g., 10-16) in order to ascertain the degree to which coordinate control requires common features shared among ribosomal protein genes of a given organism (17). In this report, we present a detailed description of the Drosophila melanogaster ribosomal protein 49 gene (11). As this is the first Drosophila ribosomal protein gene subjected to DNA sequence analysis, we are unable to compare sequence or structural features between several ribosomal protein genes of this organism. However, the rp49 gene shares many characteristics with other sequenced Drosophila genes and these are discussed below.

MATERIALS AND METHODS

Plasmid DNAs and D. melanogaster RNAs were prepared as described by Barnett et al. (18).

DNA fragments were radiolabelled at their 5'-ends with T_4 polynucleotide kinase (NEN) by the method of Maxam and Gilbert (19).

DNA sequencing was carried out using the forward-backward technique of Sief et al. (20) except that DNA polymerase I (NEN) was used at 4 units/cleavage reaction and 50 picograms DNAse I (Worthington)/cleavage reaction was added. Because the restriction maps of the plasmids were of low resolution, all sites (but one) were sequenced through from another position to avoid missing small restriction enzyme fragments not detected in the original mapping. The one exception was the Hind III site between subclones HR0.6 and H8.0. In this case, the Hind III fragments of phage c25 were 5'-end labelled and electrophoresed on a DNA acrylamide gel (not shown). No fragments smaller than the 4.0 kb fragment predicted by the original mapping were detected. The rp49 DNA sequence is presented in Figure 3B.

The S_1 nuclease protection studies were carried out according to the method of Berk and Sharp (21) or that of Sollner-Webb and Reeder (22) using sequence ladders to determine the end points of the radiolabelled mRNA/DNA hybrids. The "Blot-S_1" experiment was carried out by electrophoresing non-radioactive S_1 resistant mRNA/DNA hybrids on neutral (23) and alkaline (24) agarose gels, transferring the gels to nitrocellulose (25), and hybridizing the filters with subcloned DNA, radiolabelled by nick-translation (26).

Primer extension analysis of the rp49 intron was carried out with 250,000 cpm of 5' end labelled primer hybridized to 10 ug of adult pA^+ RNA according to the method of Klemenz and Guidescheck (27).

Denaturing RNA gels (28) were 1.5% agarose containing 2.2M formaldehyde. Samples were prepared in 1 X electrophoresis buffer (0.02M MOPS pH7.0, 0.005 M sodium acetate, 0.001 M EDTA containing 50% formamide and 2.2 M formaldehyde) and heated to 65°C for 3 min. prior to loading. Transfer of RNA to nitrocellulose paper was according to Southern (25).

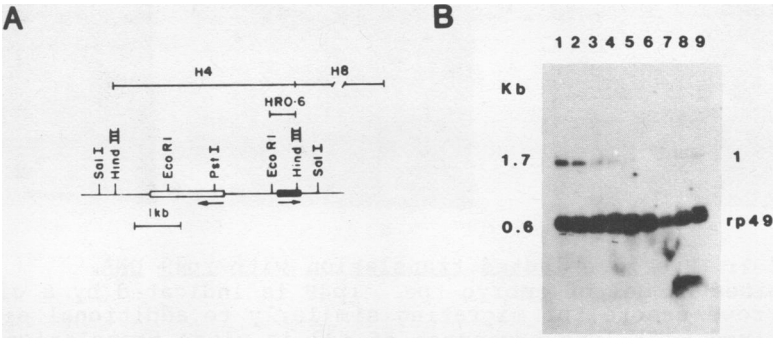


FIGURE 1: Gene expression from subclone H4.
A. Map of the region of DNA including subclone H4. The open box represents gene 1 and the closed box the rp49 gene. Arrows denote direction of transcription. The locations of subclone HR0.6 and subclone H8 are indicated.
B. 0.5 ug of pA^+ RNA, isolated from staged animals, was analyzed by a Northern blot and probed with radiolabelled subclone H4.
lane 1, 0-1 hour embryo
lane 2, 2-4 hour embryo
lane 3, 4-6 hour embryo
lane 4, 6-8 hour embryo
lane 5, 20-21 hour embryo
lane 6, 3rd instar larvae
lane 7, mid-pupa
lane 8, adult male
lane 9, adult female

Hybrid selected translation of RNA complementary to the rp 49 gene was carried out as described (11) or according to Ricchardi et al. (29). Isolation and two-dimensional gel electrophoresis of ribosomal proteins from Drosophila were carried out as described (11).

RESULTS

Location and Sequence of the rp49 gene.

The rp49 phage was selected by screening a genomic library with radiolabelled cDNA prepared from size fractionated mRNA. Positive phage were subsequently screened by hybrid selected translation. One recombinant phage, c25, selected a mRNA that encodes a 20,000 dalton large subunit ribosomal protein designated rp49 by our numbering system (11).

The rp49 mRNA is complementary to a plasmid subclone, H4, derived from phage c25. RNA blot experiments with radiolabelled H4 DNA detect two transcripts of different sizes and temporal expression (30 and Figure 1). The major 0.6 kb transcript is

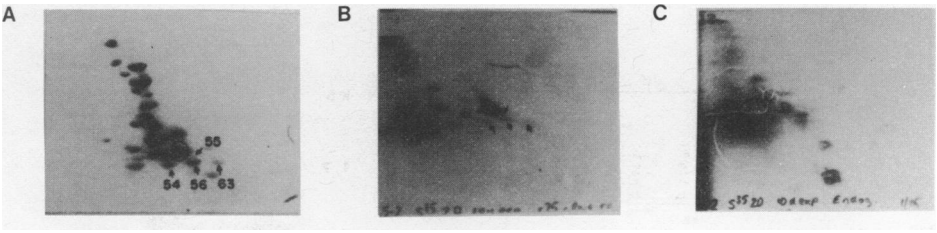


FIGURE 2: Hybrid selected translation with rp49 DNA.
A. Stained 2D gel of embryo rps. rp49 is indicated by a circle. The arrows denote rps migrating similarly to additional signals visualized with long exposures of S^{35} in vitro translation products of hybrid selected mRNAs.
B. Autoradiogram of a hybrid selection with HR0.6 DNA and 10 ug of embryo pA^+ RNA. Arrows indicate the "extra" translation products. Exposure time = 10 days.
C. Autoradiogram of a mock hybrid selection with HR0.6 DNA but without RNA. Exposure time = 10 days.

present throughout development and is the rp49 mRNA (see below). The minor 1.7 kb transcript (encoded by gene 1) is abundant in embryos and also detectable at every developmental stage examined. E.M. R-loop experiments and RNA blotting with subclones position the two genes as shown in Figure 1A (30 and data not shown). The arrows (Figure 1A) denote the direction of transcription and are consistent with S1 experiments on gene 1 (data not shown) and an extensive analysis of the rp 49 gene (see below).

By hybrid selection with subclone HR0.6, the major 0.6 kb transcript has been shown to code for rp49 (data not shown and Figure 2). Interestingly, very long exposures of the autoradiogram show, in addition to an overexposed rp49 signal, an array of fainter spots that migrate precisely with some rps (e.g., rp55 and rp 56) or very similarly to others (e.g., rp54 and rp63). This result suggests that the HR0.6 DNA has homology to the mRNA encoding these "extra" polypeptides (see Discussion). In any case, this and other hybrid selection experiments localize the rp49 gene to the HR0.6 subclone (and environs), which was subjected to DNA sequence analysis.

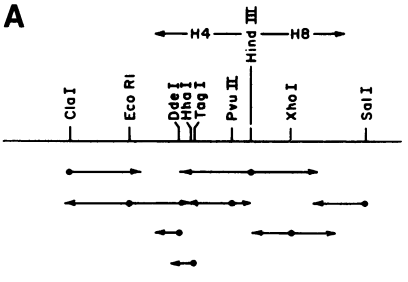
Sequence and structure of the rp 49 mRNA

Figure 3A diagrams the sequencing strategy and the restriction enzyme sites used. The rp49 DNA sequence is presented in Figure

3B. When the GC content of the rp 49 gene is plotted, the pattern shown in Figure 3C is obtained. Displayed with the graph is the structure of rp 49 mRNA (see below). Coding regions of the rp49 sequence have a significantly higher GC content than the untranslated 5' and 3' regions of the mRNA and the intron sequences. This relatively high GC content of coding DNA may be useful in distinguishing potential coding from non-coding DNA, although the generality of this observation has yet to be established.

The existence of an intron within the rp 49 gene was suggested by a preliminary analysis of the DNA sequence for open reading frames and GC content (Figure 3C). Also, several S_1 nuclease experiments (21), utilizing 5'-end labelled restriction enzyme sites within the major 3' exon, were consistent with this. mRNA, hybridized to a DNA fragment 5' labelled at the TaqI site at +187, protected a 22 bp fragment as analyzed on denaturing gels (data not shown). A larger fragment was observed on non-denaturing gels, suggesting that this fragment contained an intron. When this hybrid was run alongside the same fragment subjected to the sequencing reactions, the hybrids comigrated with the sequence 5'TCTCCTCAG3'. This sequence contains the AG sequence for a 3' splice junction site (31) and is very similar to a D. melanogaster version of this consensus sequence, YNYYYCAG (Y=Pyrimidine) (32).

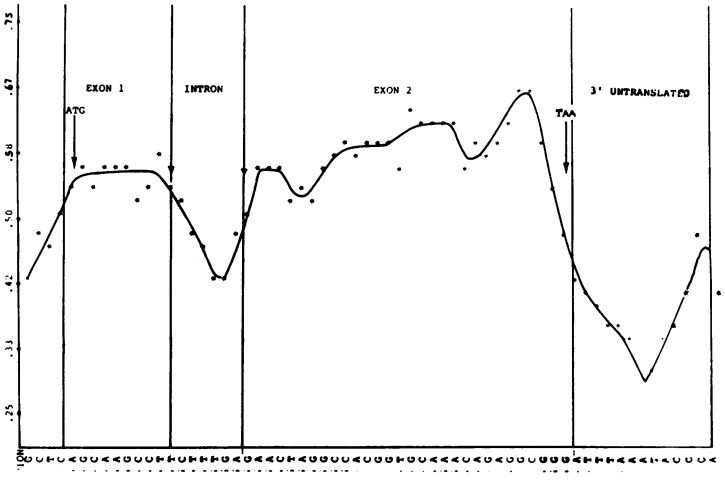
To determine the precise position of the 5' and 3' ends of the intron, a primer extension experiment (27) was carried out. A double stranded primer was derived from the +187 end labelled Taq I site by recutting this fragment with Hha I (+172)(see Figure 4A). The primer was denatured, hybridized with adult pA^+ RNA, and extended with AMV reverse transcriptase in the presence of dideoxynucleotide chain terminators to generate the sequence ladder shown in Figure 4B. This DNA sequence does not contain the 58 bp of DNA between positions +103 and +161. The genomic DNA sequence preceding the 3' junction at position +161 is as suggested by the S_1 experiments described above. The genomic sequence following the 5' junction at the G nucleotide at +102 is 5'GTGAGT3'. This sequence contains in the correct position the invariant GT of the eukaryotic 5' junction splicing



B

```
-410      -400      -390      -380      -370      -360
ACGACGTTCGATGTTTAACCACAGCTTTCTTTCGCTTCTGTTTCCGGCAAGGTATGTGCC

-350      -340      -330      -320      -310      -300
GTGATTTTGGGCCCACGTGTATGTCCATTAATTTTAAGCCGTAATGTCGTTTTTGCGTTT

-290      -280      -270      -260      -250      -240
GGAGTTGAACTGCGTTAGTCGTCGGGCTAGTGAACTAGTTAGCAAGTAGTTGCGGCTAGT

-230      -220      -210      -200      -190      -180
ATTTCAGACCATTCTTGATTGCTGTGAGCAGTTACTGCCGAATGGCTTCTGTGTTTGCTG

-170      -160      -150      -140      -130      -120
AATTCGGTATTCGATGTTCGACATCACGGTACTGTCAATGGATACTGCCCAAGCAGCTAG

-110      -100      -90       -80       -70       -60
GCCAACCTGGTTGAATTATGGATTAGTGGGACACCTTGTGTGTTATTAGCTTGATAAGTG

-50       -40       -30       -20       -10       1
ATATTTCCAGTGGGTCAGTGGACTAATGGCTACACTTGTTGTGTCCTACCAGCTTCAAG

10        20        30        40        50        60        70
ATGACCATCCGCCCAGCATAGAGGCCCAAGATCGTGAAGAAGCGCACCAAGGACTTCATCG

          80        90        100       110       120       130
GCCACCAGTGGGATCGATATGCTAAGCTGTCGCTGAGTTGCAAGGATTGTGCCAAATTGT

          140       150       160       170       180       190
ACCCGTGTTTAATCAACATGTCTCCTTGCAGCACAAATGGCGCAAGCCCAAGGGTATCGA

          200       210       220       230       240       250
CAACAGAGTGGGTCGCCGCTTCAAGGGACAGTATCTGATGCCCAACATCGGTTACGGATC

          260       270       280       290       300       310
GAACAAGCGGACCCGCCACATGCTGCCCAACGGATTCAAGAAGTTCCTGGTGCACAACGT

          320       330       340       350       360       370
GCGCGAGCTGGAGGTCCTGCTCATGCAGAACCCGCGTTTACTGCGCGAGATGCCCACGGC

          380       390       400       410       420       430
GTCTCCTCCAAGAAGCAAGAAGATTATCGAGCGCGCCAAGCAGCTGTCGGTCCGCTCACC

          440       450       460       470       480       490
AACCCCAACGGTCGCCTGCGTCTCAAGAAGAACGAGGTAAGCTTAAGATTCTTGAGAGTT

          500       510       520       530       540       550
CTTGTAACGTGGTCGGAATACACATTTGTAAACGTTAATATACCGGACTTTTAGTTAAAA

          560       570       580       590       600       610
AATGATGTGCCAGTGCCGAGTTCAATTGTCATTTCTGAGATCGGGATAGCAGCACCATCG

          620       630       640       650       660       670
ATAACATGTGCATTATCTGGATGGATATCAGTTAATCCAGACCATTGCGGTCTTTCTTTC

          680       690       700       710       720       730
TGATAGCAACTGCCTCGAGATATTAGACCAATATAAATTCTTGACGTGCAAAACTAGAC

          740       750       760       770       780       790
AGCATCAATCCTTATCAGGGAATTTTGTTATATATTTTACATTTTTCCCCCTTAGTATTC

          800       810       820       830       840       850
AAAGAGGTTGTTTATATGAAATCATATATATATTCGCAATTATTTTTACAGAACAGTGTA
```

C



consensus sequence (31) and matches closely a Drosophila consensus sequence, 5'G/GTPAGT3' (P=Purine) (32). This experiment establishes the intron boundaries with maximal precision.

The strategy for locating the ends of the gene is diagrammed in Figure 5A. To determine the 5' end of rp 49 mRNA, fragment a, 5' end labelled at the Dde I site, was incubated with embryo pA^+ RNA and digested with S_1 nuclease. The mRNA-DNA hybrids were electrophoresed on a denaturing polyacrylamide gel, adjacent to the same DNA fragment subjected to the sequencing reactions. The nibbling effect of S_1 nuclease (22) results in a ragged end on the protected DNA, although the possiblity that the 5'-end of this transcript is heterogeneous cannot be excluded. These bands comigrate with the sequence 5'ACCAGCTT3' when the inverse complement of the DNA sequence is read. Although only reminiscent of the proposed eukaryotic consensus start sequence $PyA(Py)_5$ (where A is the +1 nucleotide) (33-35), it is very similar to several D. melanogaster mRNA start sequences (32,36,37,49,50).

The most intense band in Figure 5B, lane 2, comigrates with the A at position +4 of the start sequence 5'ACCAGCTT3'; this A is therefore probably the start site of rp 49 gene transcription. The Drosophila genes for yolk protein 1 and yolk protein 2 have been shown to start at an analogous position by a

FIGURE 3A: Sequencing strategy for the rp 49 gene. The map shows the relevant restriction enzyme sites. Digested DNAs were 5'-end labelled, recut (such that only one end would be labelled), and subjected to the sequencing reactions as in Materials and Methods. The arrows indicate how far a sequence was read from a particular site.
FIGURE 3B: Nucleotide sequence of rp 49 gene. The sequence shown is the non-coding strand and corresponds to the mRNA. It is written from left to right, in the 5' to 3' direction. Underlined sequences indicate a CAT-like sequence (at -90), TATA box-like sequence (at -50), capping site (at +1), first ATG codon (at +10), intron (from +103 to +161), termination codon (at +468) and putative poly A recognition sequence (at +527). The +1 nucleotide is indicated by a heavy arrow. Lighter arrows indicate the splice junctions of the rp49 intron.
FIGURE 3C: GC content of the rp49 gene. The GC content was calculated for 25 bases to either side of every tenth nucleotide position. Arrows indicate the beginning and end of the rp49 open reading frame.

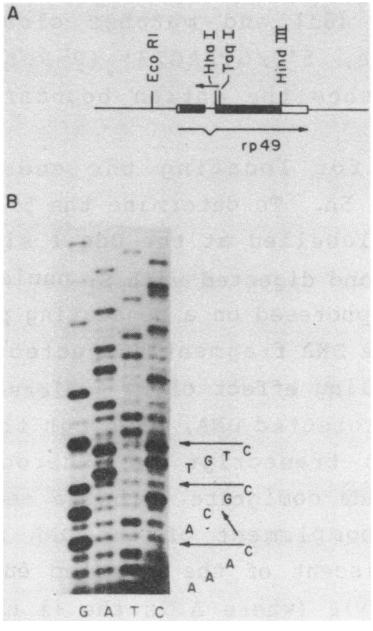


FIGURE 4: Primer extension analysis of the rp49 intron.
A. Map depicting the strategy for the primer extension experiment. The relevant restriction enzyme sites are indicated. A Taq I fragment terminating at the +187 site was 5'-end labelled and recut with Hha I to generate the radiolabelled 15 bp primer. This fragment (200,000 cpm) was hybridized with 10 ug of adult pA^+ RNA in 5X cDNA synthesis salts (15) for 45 min. at 68°C. The upper arrow indicates the direction of cDNA synthesis; the lower arrow, the direction of transcription of the rp49 mRNA.
B. The primer extension products were analysed on a 40 cm, 8% acrylamide, DNA sequencing gel. The resulting sequence ladder omits the nucleotide sequence between positions +161 and +103 when compared to the sequence determined for genomic DNA (Figure 3B). The splice junction is indicated by the arrow at 45°C. The resulting sequence ladder has been converted to that of the non-coding strand to correspond with the mRNA and Figure 3B.

similar approach (36,37). Given the heterogeneity of the protected DNA, however, it is possible that the first A of the start sequence at +1, rather than the second at +4, is the mRNA start nucleotide.

The rp49 5' upstream sequence is somewhat unusual relative to that of many other eukaryotic genes. No clearly recognizable Hogness-Goldberg box (TATA box) is present at the usual position 20-30 bp upstream from the mRNA start site (33,34). An AT rich

TABLE I. Drosophila CAAT-Like Sequences.[a]

ATGCATTAG	rp 49
AAGCAAGTC	yolk protein 1
GTGCATTAT	yolk protein 2
ATGCAAGAT	cuticle protein I
AAGCAATTC	cuticle protein II
ATGCATCAC	cuticle protein III
TTGCATCAG	cuticle protein IV
A(A/T)GCA(A/T)N(A/T)N	Drosophila consensus (from above)

[a]Shown are several Drosophila CAAT-like sequences and that of rp49. The rp49 sequence is indicated in Figure 3B. The yolk protein 1 sequence lies at approximately -70 (36). The yolk protein 2 sequence lies at approximately -65 (37). The cuticle protein sequences lie from -60 to -85 (32). The Drosophila consensus sequence is derived from these 7 sequences.

sequence, 5'GATATT3', is, however, found at approximately -50. The sequence 5'ATGCATTAG3', vaguely analogous to the CAT box or CAT consensus sequence often found 70-80 bp upstream from the mRNA start point(38), is present at position -90. Although this rp 49 sequence resembles very poorly the proposed eucaryotic CAAT consensus sequence [5'GG(C/T)CAATCT3'] (38), it is similar to one observed for several Drosophila genes as shown in Table I. Since the displacement upstream of the rp 49 CAAT-like sequence is similar to the displacement of the ATA rich sequence, the relative distance between the two putative upstream elements is similar to that found in other genes. This suggests the two regions may be functional and may play an analogous role in rp49 gene expression to the TATA box and CAT box in other genes.

The 3' end of rp 49 mRNA was determined by the following strategy. Phage c25 was hybridized with embryo polysomal pA^+ RNA and treated with S_1 nuclease. The nuclease resistant mRNA/DNA hybrids were electrophoresed on neutral and denaturing agarose gels and transferred to nitrocellulose. The filters were hybridized with radiolabelled probe b (see Figure 5A and C) to visualize rp49 DNA. On a neutral gel, the mRNA/DNA hybrid

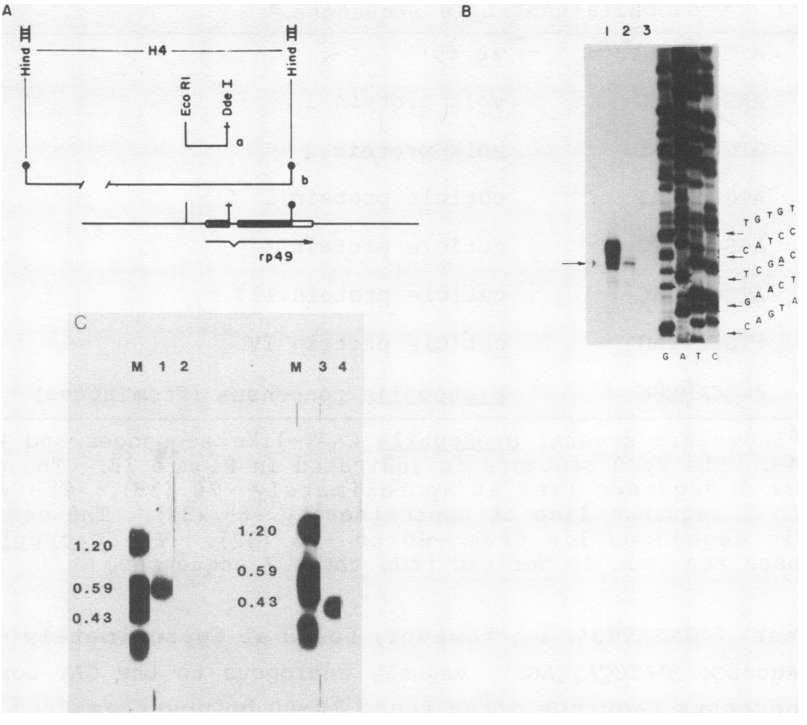


FIGURE 5: Strategy for determining the 5' and 3' ends of the rp49 transcript.
A. Restriction map indicating the relevant fragments (a and b) and regions of the rp 49 gene. The approximate location of the rp49 coding regions are shown with filled boxes, the 5' and 3' untranslated regions with open boxes. The intron is depicted by a line between the two filled boxes.
B. Fragment a (20,000 cpm), 5'-end labelled at the DdeI site, was hybridized to 10 ug of embryo polysomal pA^+ RNA and treated with S_1 nuclease as in Materials and Methods. The nuclease resistant DNA was analyzed on a 40 cm 8% DNA sequencing gel, adjacent to 5' end-labelled fragment a subjected to the forward cleavage reactions. Lane 1 was treated with 30 units of S_1 nuclease, lane 2 with 100 units of S_1 nuclease, lane 3 contains DNA incubated without of RNA. The DNA sequence ladder has been converted into the non-coding strand to correspond to the mRNA and Figure 3B. A heavy arrow marks the putative +1 nucleotide.
C. Blot hybridization analysis of the rp49 gene. 250ng of DNA from phage c25 was hybridized with 10 ug of embryo pA^+ RNA and treated with 100 units of S_1 nuclease. The S_1 resistant RNA/DNA hybrids were analyzed on native and denaturing (alkaline) agarose gels and transferred to nitrocellulose filters. Fragment b (panel A) was nick-translated and hybridized to these filters to detect the hybrids.
M, molecular weight markers. (numbers = molecular weight in kb).

lanes 1,2; analysis of S_1 resistant DNA on a neutral gel. The rp 49 mRNA/DNA hybrid has a mobility of 0.54 kb. The faint higher molecular weight hybrid probably corresponds to the 1.7 kb gene 1 transcript. lane 1, hybridization with RNA. lane 2, mock hybridization without RNA.
lanes 3,4; analysis of S_1 resistant DNA on a denaturing gel. The rp49 DNA has a mobility of 0.42 kb. lane 3, hybridization with RNA. lane 4, mock hybridization without RNA.

migrates at 0.54 kb, in accord with the mobility of rp49 mRNA on RNA gels (if an average length of the poly A tail is taken into account). On a denaturing gel, the protected DNA fragment has a mobility of 0.42 kb. The difference in mobility between the two gel systems is consistent with the presence of an intron and a 5' exon of 102 bp. The size of the 3' exon (0.42 kb) places the 3'-end of the gene at approximately +580. The sequence 5'AATATA3' is present at position +530. This is similar to the poly A addition recognition sequence, 5'AATAAA3', found upstream from the 3' terminal poly(A) tract of many eukaryotic mRNAs (39,40). Taken together with the intron assignment and the 5' end assignment, the S_1 mapping of the 3' end of the mRNA indicates a mature mRNA size of 520 bp plus the length of the poly A tail. This estimate agrees well with the mRNA size from RNA blots (~600 nucleotides) and predicts a poly A tail of 70-80 nucleotides.

Protein Coding Sequence of rp49

Eukaryotic ribosomes generally initiate protein translation at the AUG closest to the 5' end of the mRNA (41). The first AUG is present at +10 in the rp49 mRNA sequence. When the intron is removed, this open reading frame (rp49) extends 399 nucleotides to the TAA termination codon at position +468. The predicted amino acid sequence is presented in Figure 6A. The 133 amino acid polypeptide encoded by the rp49 gene is approximately 18,900 daltons in molecular weight, close to the 20,000 dalton size of rp 49 as measured by SDS gel electrophoresis (11).

Another open reading frame of 387 bp occurs in rp 49 mRNA and could encode a protein of similar molecular weight to rp49 (Figue 6B). Several arguments suggest this sequence is not the rp49 open reading frame. First, translation of this polypeptide would not begin with the first AUG (41). Second, the putative protein is much less basic than rp49 and less basic than

A

```
                                10          20          30          40
                                *           *           *           *
                               ATG ACC ATC CGC CCA GCA TAC AGG CCC AAG ATC GTG
                               Met Thr Ile Arg Pro Ala Tyr Arg Pro Lys Ile Val

     50          60            70          80          90          100 ↓
     *           *             *           *           *           *
AAG AAG CGC ACC AAG GAC TTC ATC CGC CAC CAG TCG GAT CGA TAT GCT AAG CTG TCG CAC
Lys Lys Arg Thr Lys Asp Phe Ile Arg His Gln Ser Asp Arg Tyr Ala Lys Leu Ser His

     170         180           190         200         210         220
     *           *             *           *           *           *
AAA TGG CGC AAG CCC AAG GGT ATC GAC AAC AGA GTC GGT CGC CGC TTC AAG GGA CAG TAT
Lys Trp Arg Lys Pro Lys Gly Ile Asp Asn Arg Val Gly Arg Arg Phe Lys Gly Gln Tyr

     230         240           250         260         270         280
     *           *             *           *           *           *
CTG ATG CCC AAC ATC GGT TAC GGA TCG AAC AAG CGC ACC CGC CAC ATG CTG CCC ACC GGA
Leu Met Pro Asn Ile Gly Tyr Gly Ser Asn Lys Arg Thr Arg His Met Leu Pro Thr Gly

     290         300           310         320         330         340
     *           *             *           *           *           *
TTC AAG AAG TTC CTG GTG CAC AAC GTG CGC GAG CTG GAG GTC CTG CTC ATG CAG AAC CCG
Phe Lys Lys Phe Leu Val His Asn Val Arg Glu Leu Glu Val Leu Leu Met Gln Asn Pro

     350         360           370         380         390         400
     *           *             *           *           *           *
CGT TTA CTG CGC GAG ATG CCC ACG GCG TCT CCT CCA AGA AGC AAG GAG ATT ATC GAG CGC
Arg Leu Leu Arg Glu Met Pro Thr Ala Ser Pro Pro Arg Ser Lys Glu Ile Ile Glu Arg

     410         420           430         440         450         460
     *           *             *           *           *           *
GCC AAG CAG CTG TCG GTC CGC TCA CCA ACC CCA ACG GTC GCC TGC GTC TCA AGA AGA ACG
Ala Lys Gln Leu Ser Val Arg Ser Pro Thr Pro Thr Val Ala Cys Val Ser Arg Arg Thr

     470
     *
AGG TAA
Arg ---
```

B

```
   170         180           190         200         210         220
    *           *             *           *           *           *
ATG GCG CAA GCC CAA GGG TAT CGA CAA CAG AGT CGG TCG CCG CTT CAA GGG ACA GTA TCT
Met Ala Gln Ala Gln Gly Tyr Arg Gln Gln Ser Arg Ser Pro Leu Gln Gly Thr Val Ser

   230         240           250         260         270         280
    *           *             *           *           *           *
GAT GCC CAA CAT CGG TTA CGG ATC GAA CAA GCG CAC CCG CCA CAT GCT GCC CAC CGG ATT
Asp Ala Gln His Arg Leu Arg Ile Glu Gln Ala His Pro Pro His Ala Ala His Arg Ile

   290         300           310         320         330         340
    *           *             *           *           *           *
CAA GAA GTT CCT GGT GCA CAA CGT GCG CGA GCT GGA GGT CCT GCT CAT GCA GAA CCC GCG
Gln Glu Val Pro Gly Ala Gln Arg Ala Arg Ala Gly Gly Pro Ala His Ala Glu Pro Ala

   350         360           370         380         390         400
    *           *             *           *           *           *
TTT ACT GCG CGA GAT GCC CAC GGC GTC TCC TCC AAG AAG CAA GGA GAT TAT CGA GCG CGC
Phe Thr Ala Arg Asp Ala His Gly Val Ser Ser Lys Lys Gln Gly Asp Tyr Arg Ala Arg

   410         420           430         440         450         460
    *           *             *           *           *           *
CAA GCA GCT GTC GGT CCG CTC ACC AAC CCC AAC GGT CGC CTG CGT CTC AAG AAG AAC GAG
Gln Ala Ala Val Gly Pro Leu Thr Asn Pro Asn Gly Arg Leu Arg Leu Lys Lys Asn Glu

   470         480           490         500         510         520
    *           *             *           *           *           *
GTA AGC TTA AGA TTC TTG AGA GTT CTT GTA ACG TGG TCG GAA TAC ACA TTT GTA AAC GTT
Val Ser Leu Arg Phe Leu Arg Val Leu Val Thr Trp Ser Glu Tyr Thr Phe Val Asn Val

   530         540
    *           *
AAT ATA CCG GAC TTT TAG
Asn Ile Pro Asp Phe ---
```

FIGURE 6: Translational reading frames of the rp49 DNA sequence. A. Predicted protein sequence of rp49 using the first in-frame ATG (see Figure 3B). The heavy arrow indicates the splice point between the first and second exons of this gene. The putative start ATG and termination TAA codons are underlined.

B. Predicted protein sequence of the "other" open reading frame within the rp49 gene. This putative coding region begins at the first ATG after the 3' splice junction (see Figure 3B). The "illegal" codons present in this sequence are underlined (see text).

expected from the mobility of rp49 on polyacrylamide gels. Third, a comparison with the codon usage of several D. melanogaster genes indicates that the first open reading frame rather than the second contains the codons of abundant Drosophila gene product (see below).

Sequence analysis of genes from several different organisms has revealed a species specific bias in the usage of the several degenerate codons which code for an amino acid (e.g., 42,43). For instance in yeast, abundant transcripts use a relatively restricted set of codons, which in many cases correspond to the abundant isoaccepting tRNA species present in this species (43,44). Genes that generate less abundant mRNAs use a less restricted although still biased set of codons (43). Presumably, the extreme codon bias of abundant mRNAs is due to the selective pressure to allow efficient translation of these sequences.

To begin an examination of codon bias in D. melanogaster, coding sequences of several rather abundant mRNAs were compared for codon preference. These genes are compared with a compiled set of abundant yeast mRNAs and with rp49 (Table 3). As for abundant yeast mRNAs, the codon bias is extensive in these Drosophila genes, although the frequent codons are often different from those used in yeast. Of the 61 possible sense codons, three (ATA, ACA, CGG) are not observed in any of the eight Drosophila genes examined. The codons TTA, GTA, AGT, AAA, TGT and GGG are used infrequently. It is obvious that the codon selection of the rp49 gene is qualitatively similar to these other Drosophila genes.

In contrast, the second open reading frame starting after the splice is quite different, e.g., some of the "absent" codons listed above are used at a relatively high frequency. This open reading frame contains these three rare codons (ATA, ACA, CGG) in seven positions out of the 125 codons in this reading frame: CGG is used four times, ACA twice, and ATA once (Figure 6B).

TABLE II: Codon usage in *Drosophila*.[a]

		YP1	YP2	CP1	CP2	CP3	CP4	Ac	rp49	Dm	Y
Phe	TTT	1	1	1	1	1	0	2	0	7	0
	TTC	9	13	3	3	2	3	23	4	56	29
Leu	TTA	0	0	0	0	0	1	2	1	3	2
	TTG	5	4	1	0	1	0	3	0	14	60
	CTT	1	3	1	1	1	2	2	0	9	0
	CTC	5	2	0	1	2	1	5	1	16	0
	CTA	0	0	2	1	0	0	3	0	6	4
	CTG	21	22	2	2	7	6	42	8	102	0
Ile	ATT	8	6	0	1	0	0	7	1	22	25
	ATC	12	13	7	6	8	6	49	6	101	35
	ATA	0	0	0	0	0	0	0	0	0	0
Val	GTT	8	7	3	5	0	1	2	0	26	64
	GTC	9	8	6	3	8	9	13	5	56	44
	GTA	0	0	0	2	0	1	6	0	9	0
	GTG	14	13	2	3	3	2	23	3	60	0
Ser	TCT	1	3	0	1	2	1	2	1	10	38
	TCC	8	13	6	4	3	3	27	0	64	33
	TCA	0	0	1	2	0	0	2	2	5	0
	TCG	3	3	2	2	0	0	15	4	25	0
	AGT	0	1	0	0	1	0	0	0	2	0
	AGC	14	9	2	2	3	4	3	1	37	0
Pro	CCT	2	2	2	2	2	2	3	1	15	3
	CCC	18	18	7	6	5	5	23	5	82	1
	CCA	1	1	4	4	1	2	7	3	20	32
	CCG	1	1	1	1	0	0	5	1	9	0
Thr	ACT	0	3	1	1	1	1	3	0	10	27
	ACC	22	25	1	2	1	1	42	5	94	34
	ACA	0	0	0	0	0	0	0	0	0	0
	ACG	0	2	0	0	0	0	4	3	6	0
Ala	GCT	7	5	0	3	5	3	11	1	34	68
	GCC	26	20	12	12	8	9	39	2	126	32
	GCA	2	0	2	1	0	0	3	1	8	0
	GCG	3	1	1	1	0	0	6	1	12	0

		YP1	YP2	CP1	CP2	CP3	CP4	Ac	rp49	Dm	Y
Tyr	TAT	5	2	0	0	0	0	9	2	16	0
	TAC	10	15	2	2	3	3	22	2	57	33
His	CAT	1	3	3	2	1	0	2	0	12	1
	CAC	10	5	7	7	1	2	16	4	48	26
Gln	CAA	2	2	0	0	0	1	4	0	9	20
	CAG	33	37	2	2	3	3	20	4	100	0
Asn	AAT	5	5	1	1	1	1	2	2	16	0
	AAC	25	23	5	4	6	5	16	5	84	37
Lys	AAA	0	1	0	0	0	0	2	1	3	7
	AAG	26	20	2	2	6	6	36	13	98	69
Asp	GAT	11	10	7	4	2	3	13	1	50	17
	GAC	13	9	0	4	6	4	32	2	68	44
Glu	GAA	3	3	0	1	0	3	4	0	14	49
	GAG	21	29	9	7	6	6	50	5	128	0
Cys	TGT	0	0	0	0	1	0	2	0	3	12
	TGC	1	2	1	0	0	1	12	1	17	0
Arg	CGT	9	10	1	1	0	1	12	1	34	0
	CGC	8	13	1	2	0	0	22	12	46	0
	CGA	0	2	1	0	1	0	1	1	5	0
	CGG	0	0	0	0	0	0	0	0	0	0
	AGA	1	0	0	0	0	0	0	4	1	29
	AGG	4	7	0	0	0	0	1	2	12	0
Gly	GGT	11	7	1	1	1	1	26	3	48	90
	GGC	13	18	6	4	2	2	29	0	74	3
	GGA	7	10	4	5	4	4	3	3	37	0
	GGG	0	1	0	0	0	0	0	0	1	0

[a]Codon usage for various genes are compared. YP1 is the Drosophila yolk protein 1 gene (36). YP2 is the Drosophila yolk protein 2 gene (37). CP1-CP4 are the 4 Drosophila cuticle protein genes (32). Ac is 2 Drosophila actin genes, the codons of which have been pooled together (51). rp49 is from this paper. Dm is a sum of the preceeding columns, i.e., all 7 Drosophila columns except rp49. Y is the sum of the codon usage of 3 abundant yeast genes (2 genes coding for glyceraldehyde-3-phosphate dehydrogenase and 1 gene coding for alcohol dehydrogenase, ADH-I) from reference 43.

Thus, it seems likely that the first open reading frame is translated and less likely (by these criteria) that this "other" rp49 translation reading frame is translated, at least into an abundant gene product. One might speculate, however, that since the first AUG rule and the codon information are based on genes with reasonably abundant transcripts, this second translational frame may code for an as yet uncharacterized product that exploits an alternate, low level use of sequences in the "parent" rp49 transcription unit.

DISCUSSION

Although the rp49 gene has some unusual features, it shares most of its characteristics with other Drosophila genes. The 5' end of the gene has possible TATA box-like and CAT box-like sequences, and an extremely short (9bp) 5' untranslated region. The presence of a single, short intron is quite common among Drosophila genes (e.g., 32,36,37). The relatively high GC content of coding DNA and the putative signals at the cap site, splice junctions, and upstream of the poly A addition site, are in accord with those found in other Drosophila genes.

Very long exposures of hybrid selected translation products show an array of faint spots (in addition to the extremely intense rp49 spot) that migrate either precisely like or very similarly to other rps. At the time of the initial report of the isolation of c25, the amount of mRNA utilized and exposures were not sufficient to detect these relatively faint "extra" spots (11). With longer exposures and 10 ug pA^+ RNA, the result reported here was reproducibly obtained and independent of the source of rp49 DNA (i.e., phage c25 or subclones H4 and HR0.6) and independent of the hybrid selection technique (11,29). The result suggests that the c25 sequences may have some homology to the mRNA encoding those "extra" polypeptides.

Several observations suggest that this homology is quite limited. First, the signal of the "extra" spots is much less intense than the rp49 signal and is only visible when the rp49 spot is grossly overexposed. Blot hybridizations of radioactive rp49 DNA to electrophoretically separated DNA fragments or mRNAs detect no additional signals, although these blots generally undergo a more stringent washing procedure than the hybrid

selections. It should also be emphasized that these faint polypeptide spots are not proven to be bona fide rps. The comigration for some of the fainter spots is not perfect. In addition no attempt has been made to determine whether any of these faint "extra" spots are bona fide rps by partial peptide analysis.

A recent paper, reporting the isolation of another Drosophila rp gene (45), demonstrates that the phage DNA efficiently hybrid selects 6 rp mRNAs, although the phage contains the gene for only one of the selected rps. Bozzoni et al. (46) have also observed a similar phenomenon in their hybrid selection analysis of the Xenopus L1 and L14 rp genes. These investigators have determined that the sequences responsible for the appearance of the "extra" faint spots, which migrate similarly to other ribosomal proteins, reside very close to or within these identified genes. Since the rp49 subclone, HR0.6, selects mRNA which generates the same array of faint polypepetides as the c25 phage DNA and since the "extra" spots are faint as compared to the "real" spot, it appears that we are observing a similar phenomenon with the Drosophila rp49 gene to the one previously reported by the Xenopus group (46).

An interesting feature of the rp49 gene is the presence of an alternative uninterrupted open reading frame in addition to the putative rp49 coding sequence. Although the codon analysis data presented above suggest that this additional open reading frame is not translated and therefore is probably a random occurrence (perhaps in part a consequence of the relatively high GC content where stop codons are less likely to occur), mRNA translated in this frame could probably explain at least one of the "extra" spots in the in vitro translation products of mRNA hybrid selected by rp49 DNA.

The generality of some of the observations and comparisons reported here, in particular the GC content analysis and codon frequency, remains to be established. It will be interesting to learn how well other Drosophila genes, especially non-abundant genes, conform to the codon rules. The availability of DNA-mediated transformation (47,48) in Drosophila will allow us to test the functional requirements for upstream sequence elements.

Finally, it will be interesting to see if sequence and/or structural homologies exist between rp49 and other Drosophila ribosomal protein genes as is the case for yeast (17).

ACKNOWLEDGEMENTS

We thank M.C. Hung, H. Colot, A. Vincent, and M. Wormington for helpful suggestions and their comments on this manuscript. We are also greatful to T. Tishman for secretarial help and N. Abovich for photographic assistance. This work was supported by a grant from the NIH (GM23549).

REFERENCES

1. Warner, J.R., Tushinski, R.J. and Wejksnora, P.J. (1980) in Ribosomes, Structure, Functions, and Genetics, G. Chambliss, G.R. Craven, J. Davies, K. Davis, L. Kahan, and M. Nomura Eds., pp. 889-902, Baltimore University Park Press, Baltimore.
2. Santon, J.B., and Pellegrini, M. (1980) Proc. Natl. Acad. Sci. 77, 5649-5653.
3. Geyer, P.K. Meyuhas, O., Perry, R.P., and Johnson, L.F. (1982) Mol. Cell. Biol. 2, 685-693.
4. Pierandrei-Amaldi, P., Campion, N., Beccari, E., Bozzoni, I., and Amaldi, F. (1982) Cell 30,163-171.
5. Faliks, D., and Meyuhas, O. (1982) Nucl. Acids Res. 10, 789-801.
6. Weiss, Y.C., Vaslet, C.A., and Rosbash, M. (1981) Devel. Biol. 87, 330-339.
7. Pearson, N.J., and Haber, J. (1980) J. Bacteriol. 143, 1411.
8. Pearson, N.J., Fried, H.M., and Warner, J.R., 1982. Cell 29,347-355.
9. Kim, C.H., and Warner, J.R. (1983) J. Mol. Biol. 165, 79-89.
10. Woolford, J.L. Jr., Hereford, L.M., and Rosbash, M. (1979) Cell 18, 1247-1259.
11. Vaslet, C.A., O'Connell, P., Izquierdo, M., and Rosbash, M. (1980) Nature 285, 674-676.
12. Fried, H.M., Pearson, N.J., Kim, C.H., and Warner, J.R. (1981) J. Biol. Chem. 256, 10176-10183.
13. Bollen, G.H.P.M., Cohen, L.H., Mager, W.H., Klaassen, A.W., and Planta, R.J. (1981) Gene 14, 279-287.
14. Monk, R.J., Meyuhas, O., and Perry, R.P. (1981) Cell 24,301-306.
15. D'Eustachio, P., Meyuhas, O., Ruddle, F., and Perry, R.P. (1981) Cell 24, 307-312.
16. Bozzoni, I., Beccari, E., Luo, X.Z., Amaldi, F., Pierandrei-Amaldi, P., and Campioni, N. (1981) Nucl. Acids Res. 9, 1069-1085.
17. Teem, J.T., manuscript in preparation.
18. Barnett, T., Pachl, C., Gergen, J.P., and Wensink, P.C. (1980) J. Bacteriol. 143, 1411.
19. Maxam, A., and Gilbert, W. (1980) Methods in Enzymology 65, 499-559.
20. Seif, I., Khoury, G., and Dahr, R. (1980) Nucl. Acids Res. 8, 2225-2240.

21. Berk, A.J., and Sharp, P. (1977) Cell 12, 721-732.
22. Sollner-Webb, B., and Reeder, R.H. (1979) Cell 18, 485-499.
23. Helling, R.B., Goodman, H.M., and Boyer, H.W. (1974) J. Virol. 14, 1235-1244.
24. McDonell, M.W., Simon, M.N., and Studier, F.W. (1977) J. Mol. Biol. 110, 119-146.
25. Southern, E.M. (1975) J. Mol. Biol. 98, 503-518.
26. Rigby, P.W.J., Dieckmann, M., Rhodes, C., and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
27. Klemenz, R. and Guideschek, E.P. (1980) Nucl. Acids Res. 8, 2679-2689.
28. Rave, N., Crkvenjakov, R., and Boedtker, H. (1979) Nucl. Acids Res. 6, 3559-3567.
29. Ricchardi, R.P., Miller, J.S., and Roberts, B.E. (1979) Proc. Nat. Acad. Sci. USA 76, 4927-4931.
30. Wong, Y-C., O'Connell, P., Rosbash, M., and Elgin, S.C.R. (1981) Nuc. Acids Res. 9, 6749-6762.
31. Mount, S.M. (1982) Nucl. Acids Res. 10, 495-472.
32. Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J., and Davidson, N. (1982) Cell 29, 1027-1040.
33. Busslinger, M., Protmann, R., Irminger, J.C., and Bernstiel, M.L. (1980) Nucl. Acids Res. 8, 957-977.
34. Corden, J., Wasylyk, B., Buchwalder, A., Sarsone-Corsi, P., and Vedinger, L. (1980) Science 209, 1406-1414.
35. Breathnach, R., and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
36. Hung, M-C. and Wensink, P.C. (1981) Nucl. Acids Res. 9, 6407-6419.
37. Hung, M-C. and Wensink, P.C. (1983) J. Mol. Biol. 164, 481-492.
38. Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. (1980) Nucl. Acids Res. 8, 127-142.
39. Proudfoot, N.J., and Brownlee, G.G. (1976) Nature 263, 211-214.
40. Fitzgerald, M., and Shenk, T. (1981) Cell 24, 251-260.
41. Kozak, M. (1984) Nucl. Acids Res. 12, 857-872.
42. Grantham, R., Gauthien, C., Gouy, M., Mecier, R., and Pave, A. (1980) Nucl. Acids Res. 8, 49-62.
43. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.
44. Guthrie, C. and Abelson, J. (1982) in The Molecular Biology of the Yeast Saccharomyces. Metabolism and Gene Expression, J.N. Strathern, E.W. Jones, J.R. Broach (eds.), Cold Spring Harbor Laboratory, pp. 487-528.
45. Fabijanski, S. and Pellegrini, M. (1982) Gene 18, 267-276.
46. Bozzoni, I., Tognoni, A., Pierandrei-Amaldi, P., Beccari, E., Buongiorno-Nardelli, M., Amaldi, F. (1982) J. Mol. Biol. 161, 353-371.
47. Spradling, A.C., and Rubin, G.M. (1982) Science 218, 341-347.
48. Rubin, G.M., and Spradling, A.C. (1982) Science 218, 348-353.
49. Ingolia, T.D. and Craig, E.A. (1981) Nucl. Acids Res. 9, 1627-1642.
50. Holmgren, R., Corces, V., Morimoto, R., Blackman, R., Messelson, M. (1981) Proc. Natl. Acad. Sci. USA 78, 3775-3778.