The EMBO Journal vol.6 no.13 pp.4095–4104, 1987

Developmental expression of a regulatory gene is programmed at the level of splicing

Tze-Bin Chou, Zuzana Zachar and Paul M.Bingham

Department of Biochemistry, State University of New York, Stony Brook, NY 11794, USA

Communicated by V.Pirrotta

We report sequence and transcript structures for a 6191-base chromosomal segment containing the presumptive regulatory gene from *Drosophila, suppressor-of-white-apricot* [*su(w^a)*]. Our results indicate that *su(w^a)* expression is controlled by regulating occurrence of specific splices. Seven introns are removed from the *su(w^a)* primary transcript during precellular blastoderm development. The sequence of this mature RNA indicates that it is a conventional messenger RNA. In contrast, after cellular blastoderm the first two of these introns cease to be efficiently removed. The mature RNAs resulting from this failure to remove the first two introns have structures quite unexpected of mRNAs. We propose that postcellular blastoderm *su(w^a)* expression is repressed by preventing splices necessary to produce a functional mRNA. Implications and mechanisms are discussed.

Key words: regulation/splicing/*Drosophila/suppressor-of-white-apricot*

Introduction

Retrotransposons are developmentally programmed, somatically transcribed polymerase II transcription units. This developmental programming apparently involves parasitism of host regulatory information (see, for example, Varmus, 1983; Winston *et al.*, 1984; Zachar *et al.*, 1985; Parkhurst and Corces, 1987 for recent discussions). Retrotransposon parasites are thus attractive cases for analysis of the regulatory machinery of their cellular hosts.

In practice, retrotransposon transcription units have an advantage over host genes for such studies. In *Drosophila*, retrotransposon insertion causes a large fraction of spontaneous mutations in host genes (Zachar and Bingham, 1982; Bender *et al.*, 1983). A subset of these insertion alleles couple expression of the host gene and the inserted retrotransposon so as to generate convenient 'reporters' allowing genetic screens for host genes regulating retrotransposon expression (see, for example, Bender *et al.*, 1983; Zachar *et al.*, 1985; Parkhurst and Corces, 1986). Among host genes so identified are allele-specific suppressors and enhancers that interact quite specifically with individual retrotransposon families (Bender *et al.*, 1983; Modolell *et al.*, 1983; Levis *et al.*, 1984; Zachar *et al.*, 1985; Parkhurst and Corces, 1986).

The strict allele-specificity of suppressors and enhancers is informative. It argues that they do not influence phenotypes for generalized or trivial reasons. Rather, this specificity indicates that suppressors and enhancers are likely to be involved in processes idiosyncratic to individual retrotransposons. These, presumably, include developmental programing of expression.

Motivated by these considerations, we have studied one of these loci in *Drosophila, suppressor-of-white-apricot* [*su(w^a)*]. *su(w^a)* specifically suppresses the hypomorphic mutant phenotype of the *white-apricot* [w^a] allele resulting from insertion of the *copia* retrotransposon into the second intron of the *white* locus. This *copia* insertion contains a transcript terminus formation site at which ~95% of w^a transcripts are terminated; some or all of the remaining ~5% of w^a transcripts are processed at the conventional *white* splice sequences to produce wild-type mature *white* message accounting for the low level of *white*$^+$ function of the allele (Pirrotta and Brockl, 1984; Levis *et al.*, 1984; see Zachar *et al.*, 1985 for detailed discussion). Thus, w^a is a reporter for at least two RNA processing events: splicing of the second w^a intron and polyadenylated terminus formation in the *copia* transposon. Mutational inactivation of *su(w^a)* increases the levels of wild-type transcripts from w^a, presumably by influencing one or both of these RNA processing events (Zachar *et al.*, 1985).

Results reported here indicate the expression of *su(w^a)* is controlled at the level of splicing. Results reported in the accompanying paper (Zachar *et al.*, 1987a) indicate that this post-transcriptional regulation constitutes autoregulation. Thus, *su(w^a)*—which apparently regulates a w^a RNA processing event in *trans*—apparently regulates its own expression at the level of RNA processing.

Results

DNA sequence analysis of su(w^a)

We have shown previously by gene transfer that all DNA sequences necessary to confer a *su(w^a)*$^+$ eye color phenotype on w^a individuals are contained in a ~6.2-kb *Sal*I–*Nru*I segment from the cloned *su(w^a)* region (Zachar *et al.*, 1987b). The sequence of this 6191-base interval and a standard coordinate system are given in Figure 1.

Northern analysis of developmental changes in su(w^a) transcript pattern

Northern analysis of the developmental changes in transcript pattern from *su(w^a)* is shown in Figure 2. *su(w^a)* produces predominantly a 3.5-kb RNA in precellular blastoderm embryos. (A minor transcript at 3.7 kb is also seen at variable levels during this developmental stage.) During subsequent development, *su(w^a)* produces predominantly a mixture of 5.2- and 4.4-kb RNAs with trace amounts of the 3.5-kb RNA (Figure 2; results below).

The precellular blastoderm pattern exists briefly (for the first ~3 h of development) and is rapidly replaced by the postcellular blastoderm pattern (replacement complete by 6–8 h postoviposition; Figure 2 and results not shown). The postcellular blastoderm pattern persists throughout the rest of development with only modest quantitative variation (Figure 2).

S_1 protection analysis of su(w^a) transcript structure

Our approach exploits the availability of nested deletion sets made in M13-cloned segments of the gene in the process of DNA sequence determinaton (Materials and methods). S_1 protection analysis using these clone sets allows straightforward deletion mapping of exons. Further, it allows measurement of distances between sequenced deletion breaks and exon boundaries by siz-

```
        10         20         30         40         50          60         70         80         90        100         110        120        130        140        150         160        170        180        190        200
GTCGACTGGG ACATTATTTA CTCCAAGCTG CGAACCCTGA AGTTCCCAAA  TGCCAAGGAT TGACTGAGGG AATGTTGTGA GATTTAGTTG AAATGTAACT  GTGAATTAAA TGGTTTTTTT TTACACATGG CACAGCGAAT AAATTTTGGG  AGGTGGATAA ATATTGCAAA TGAATTTCAC TTGAAACGTG TTTGAATCCT
       210        220        230        240        250         260        270        280        290        300         310        320        330        340        350         360        370        380        390        400
ACGATGAACG GCGCCGATAC TATCGATACA AGCCCAGTGG TCGAGCACTT  CTCGATACTG CCGTGGTTGA GCAACTCCGC TCTGGCATCG CGATAACATC  GATATAACTG CCACCTCTAG CATAATATAT ATGTACTACA TTTATACAAA  TATAAATATA ATGGAAATAT GAGAATTTAA CCAAACACCG CCCGAAATAT
       410        420        430        440        450         460        470        480        490        500         510        520        530        540        550         560        570        580        590        600
GAGATCCTCC GAGCACAGAT CCTAGCAATC GATCTCCTCA TAAAGTGAGT  GAGTCGCAAG AGACGAAGGG AGCGAGAGAG AGAAAGGAGC ACTTAAATTG  CACATTCGCT GCTGTTTGTC CGCCGCGATG CTGCCCTACA ATGTGCGAAA  CGCCGGTGGC GGAAGCGTTG GCGGAATCCT GCGGCGCACG GGCCAAGGTT
       610        620        630        640        650         660        670        680        690        700         710        720        730        740        750         760        770        780        790        800
CGGGCACCGG AAGCACGATT CTTGGCAATG GAAACAGTCC GGGAGCACTA  GGAGCCGGAA AAGTGTCCTC CTCCAGCCTG GAAAACCACA GACAACCGCC  GTTGGAACTC CTCGTGTTTG GCTACGCCTG CAAGATTTTC CGCGATGACG  AGAAGGCGCG GGAAATGGAC CACGGCAAGC AGTTGATTCC CTGGATGGGC
       810        820        830        840        850         860        870        880        890        900         910        920        930        940        950         960        970        980        990       1000
GATGTAAATC TCAAGATTGA CAGGTTAATA ATTGGGACTA ACCCGAGCAA  TAATTATACG ATTAATCAGT ACGCACACCA CTGCCACAAA CACACACACA  CAGCCACATA CCCACCAATG TCGTGTCCCG TTAGGACCAC ACTCCCACAC  ACAGCAGCGG TCAAATTGTT AGCAGAGTCA GATACATACA TATATGTAGG
      1010       1020       1030       1040       1050        1060       1070       1080       1090       1100        1110       1120       1130       1140       1150        1160       1170       1180       1190       1200
TATTGTGATC GATGGTATTG TGATTATGGC GTTAGAAGAC CGAGCTTGTC  GCTATTTATA GCCCTATTTT GCATAATCTA AACAGAACTG AATTAATTTC  TATTTGATAA AATCGAGAAG AGGATTTTCT GAAAAGGATC AAACTGATCG  TCTGAGTAGA TATTCTAACA TTTGTTGTAA TAATTCAAGA CTGCGAACTG
      1210       1220       1230       1240       1250        1260       1270       1280       1290       1300        1310       1320       1330       1340       1350        1360       1370       1380       1390       1400
CAGACACACA CCCATCTATT CATCCGCCCA CACAACCAAA ACACACACGC  ACACACACCA TTGTAATAAC AGCTCTTTTC AGTGCTTCTC AAACTGTAAT  CTCTATCCAT ATACAATTTG AAAACTGTAA TTCTTTGCCC ACTTTAAATG  TAATCCGATT AACTTTTTAG CCACCGGAGC GGTATGTATA CGAACGATCT
      1410       1420       1430       1440       1450        1460       1470       1480       1490       1500        1510       1520       1530       1540       1550        1560       1570       1580       1590       1600
CTAGCCTTAA AACCTAATCT TATCTGAAAC TGCATTATAT ATACTATATA  CCGTATGACT ATCAATATTG TGTAAATGCG CGAGAGTGTG CGAAGCAAAT  GGGACGATTT TCAAGGAGCG GCCCTGCCTC CCGTTTGTTT CGCCCGTCCC  GTTTCGCGTC GCCCCCCAAA ATTGCCACAC CCACTGACAA TGCTTCTCCC
      1610       1620       1630       1640       1650        1660       1670       1680       1690       1700        1710       1720       1730       1740       1750        1760       1770       1780       1790       1800
TCCTCTCTGT TTATATTTCT CTGCTTGCTT TCTTTCGGCT CCGCAAACAC  CATTGCCATT GCGATTACCG TGACCACAGA TACGATGTGC GTGGAGCCCT  CTGCGAACTG GCACCGCACG AGGCGCCACC AGGCGGATAC GGTAACCGGC  TGGAGTACTT GAGTGCCGAG GAGCAGCGCG CCGAGCAGCT GTGCGAGGAG
      1810       1820       1830       1840       1850        1860       1870       1880       1890       1900        1910       1920       1930       1940       1950        1960       1970       1980       1990       2000
GAGCGGTACC TGTTTCTGTA CAATAACGAG GAGGAGCTGC GACTGCGACA  AGGTGAGCAC CACGGTGGTC TGGTAGTACC TCCAGATCCA AAGAACCGGG  CAGGGGTGTC GATATACTTT ACTTGGTTTC GAAACTCCAG CAATCTCTTC  TTCGAATGTA AAGACGAAGA GAAGAGATTT TTGGAGCATA GGGAAAATGT
      2010       2020       2030       2040       2050        2060       2070       2080       2090       2100        2110       2120       2130       2140       2150        2160       2170       2180       2190       2200
CGGTTTTAAT ACTCTAGGCT AAAAGCTTTA AAAACTAGTT AATGTCAAAT  TAAGTTGGGC TGGGACCTAC TCACATCAGA TTTCTAACGC ATGTAAACGT  ATCGACACTC CCAATCCCGA ATTCCTCAAA TTGTAAATTG TATTACAAGT  ACTGCAAGTG TCATGAAACG CAAAACGCAA TCAAACCCAA ATCGCAAATC
      2210       2220       2230       2240       2250        2260       2270       2280       2290       2300        2310       2320       2330       2340       2350        2360       2370       2380       2390       2400
GAAAATAACT TGAACAAATC AGTACCAATC ACATCAAATC CAACCATCGA  AAAAGATACC ACCGACATCC CCTCGATATT TCCAGCGAGT CCGTTCTTGG  AGCAGGTAAA TACGTAATCC TCACTACCTG ACAGAATAAA TCGTTTCTTT  CAGGTTGTAT TAGCTGGAGC GCCTTCTTGG GCGTGTGTTG TCAGGGATCC
      2410       2420       2430       2440       2450        2460       2470       2480       2490       2500        2510       2520       2530       2540       2550        2560       2570       2580       2590       2600
AAATGCATGA TCAGCTTGCC GGGGAAGGCT CAGGATCAAG CGATTTTTGA  TATCCAGCAC CTTCCTCTAT CTCTTTCTCT CTGTATCTCT CTCTTGTTCG  CATCAACTCG AATACGTTTC TCTCGCTCTG ATTCGAATTA GTTATTGCTG  CCACCCCTCT CCTCCCGTAC TAAAATCGCA CATCTTCGTG TGTTCTCTGC
      2610       2620       2630       2640       2650        2660       2670       2680       2690       2700        2710       2720       2730       2740       2750        2760       2770       2780       2790       2800
TCTTATTTTG GCCCAATTCG CTGATCCTGC AGAGGAGGAC CTCAAGCGAC  TGCAGCAGGA AACCTCGGGC GGCTGCTTCA GCCAGGTGGG CTTTCAGTAC  GATGGCCAGT CGGCAGCCTC CACTTCCATC GGTGGCTCCT CAACGGCGAC  CTCCCAGCTC TCGCCGAACT CCGAGGAGAG CGAGCTGCCT TTCGTCCTTC
      2810       2820       2830       2840       2850        2860       2870       2880       2890       2900        2910       2920       2930       2940       2950        2960       2970       2980       2990       3000
CCTACACCCT GATGATGGCA CCACCCTTGG ACATGCAGCT GGTGAGTCAA  AGACTAAATC ACTTCACACC AAAAAGTTAT GGTAGTCATG CTTTCTTAAA  TGCTTTCCAG CCCGAAACGA TGAAGCAGCA CGCTATCATA GAGAAGACTG  CTCGCTTTAT CGCAACCCAG GCCGCCCAAA TGGAAATTCT TATCAAGGCA
      3010       3020       3030       3040       3050        3060       3070       3080       3090       3100        3110       3120       3130       3140       3150        3160       3170       3180       3190       3209
AAGCAGGCGA ACAACACGCA GTTCGATTTC CTGACCCAAG GTGGACATCT  GCAGCCTTAC TACCGTCACC TTTTAGCTGC TATCAAGGCG GCCAAGTTCC  CGCCCGCACC ACAAACTCCG TTGGATCAGC AGAATACAGA CAAGGAAGCA  CCCTCAGCGG ACGACCACAG CGAGGAAGTC GCGGGTGGAA GGAGGAACCC
      3210       3220       3230       3240       3250        3260       3270       3280       3290       3300        3310       3320       3330       3340       3350        3360       3370       3380       3390       3400
CAACCAAGTG GTCATCACGG TGCCCACCAT CAAGTACAAG CCCTCGGCCA  ACTGCGCCTA CACCCAGCTA ATTAGCAAGA TCAAGGGCGT GCCCCTCCAA  GCGGTCTTGC AGGAGGACGA GTCAAGCAAT CCGGGCAACT CGCAGCACTC  GGGCGGCACG GCCTCGCCCG CACTTAGCTG TCGATCCGAA GGTCACAACA
      3410       3420       3430       3440       3450        3460       3470       3480       3490       3500        3510       3520       3530       3540       3550        3560       3570       3580       3590       3600
GCCAGGGAGG GGAGTTCACA CCCGTATTGC TGCAGTACAA CGGGAGCACC  TTCACACACG AGGAGGAATC GTCGAACCGG GAGCAGCAGG ATGATAACGA  TGTCAACGGT GGCGAGCCGC CCCAGGTGGA GCTGCTTAAG AACACGAGTG  CGCTGGCGTT GGCCCAAAAC TATAGTTCAG AAAGTGAGGA GGAGGAAGAT
      3610       3620       3630       3640       3650        3660       3670       3680       3690       3700        3710       3720       3730       3740       3750        3760       3770       3780       3790       3800
CAAGTTCAGC CAGAAAAGGA AGAGGAAAAG GTGGGCACTC TCACAATATA  AAGTTCATGT GTATAATCTG TCATTTCTTT ATTTTAGAAA CCAGAGCCCG  TGCTTACCTT CCCCGTTCCG AAGGACAGTC TAAGGCACAT AATCGATAAG  ACGGCCACGT ACGTAATCAA GAACGGACGC CAATTCGAGG AGACCCTTCG
      3810       3820       3830       3840       3850        3860       3870       3880       3890       3900        3910       3920       3930       3940       3950        3960       3970       3980       3990       4000
CACGAAGAGC GTGGATCGGT TTAGTTTCTT GCTGCCGGCC AATGAGTACT  ATCCATACTA CTTGTACAAA GTGACCGGCG ATGTGGATGC GGCTTCCAAG  GAGGAGAAAA CTAGGAAAGC AGCAGCCGTA GCTGCCGCCC TGATGTCAAA  GAAGGGTCTT AGTTTTGGAG GAGCAGCAGC TGCAGTGAGT GGAAGCAATC
      4010       4020       4030       4040       4050        4060       4070       4080       4090       4100        4110       4120       4130       4140       4150        4160       4170       4180       4190       4200
TGGATAAGGG TGAGTTAAGG ACAAAGGGAT TTCCCGGGTT CGGTATTTAA  TTTGTATTTG TTTTTGTAGC TCCTGTAAGT TTCTCCATTC GGGCTCGCGA  TGACCAATGT CCATTGCAAC ACACTTTACC GCAGGAGGCC AGCGACGAGG  AGACTAGTTC GAATGCTGCT GGGGTGGAGC ATGTGCGACC AGGTATGCCA
      4210       4220       4230       4240       4250        4260       4270       4280       4290       4300        4310       4320       4330       4340       4350        4360       4370       4380       4390       4400
GATAGTGTGC AACGGGCCAT CAAACAGGTT GAGACGCAGT TGTTGGCCAG  GACCGCCGGG CAGAAGGGAA ACATCACGGC TTCCCCATCC TGTTCCTCAC  CGCAAAAAGA GCAGCGGCAA GGTACGATAT AGTTATATAT ATATTAAGTC  CACTCTGATA TTAATAATAG CACAATTTTG TCGTGAAGCT GAGGAGCGTG
      4410       4420       4430       4440       4450        4460       4470       4480       4490       4500        4510       4520       4530       4540       4550        4560       4570       4580       4590       4600
TTAAGGATAA GCTGGCGCAG ATTGCCCGCG AGAAGCTGAA CGGAATGATT  TCACGCGAAA AACAATTGCA GTTGGAACGA AAGCGAAAGG CCTTGGCCTT  TCTAAATCAG ATAAAAGGTG AGCTTACCGT AAATCGAACG CTATTGTGTA  TAATCTCGTC GATTTCTACC AGGTGAGGGT GCCATTGTGG GCTCTGCTGT
      4610       4620       4630       4640       4650        4660       4670       4680       4690       4700        4710       4720       4730       4740       4750        4760       4770       4780       4790       4800
GCCTGTTGTT GGTCCGAATC CTCCAGAGTC CGCGGCCGGA GCTGCGACCG  CCGACAGTGG TGACGAGTCT GGTGATTCTG TACGTTCCAT ACCAATTACC  TACTTTGGCC CCGACGACGA TGACGAAGTC GGTGAACAGA GGCCTGAGAT  GAGGCTCATA GGGAGCACGC AGAAGGACGA GGAGGACGAC GATGAAGAGG
      4810       4820       4830       4840       4850        4860       4870       4880       4890       4900        4910       4920       4930       4940       4950        4960       4970       4980       4990       5000
ATGGCGGCGA TCTGGAGAAG TACAACTTGC TCAACGACGA TAGCACGAAC  ACCTTCACCA GCAAACCCGT GCTTCCTCCC ACCGCTGCTC CACCTCCTGC  TGCTGTTCTG CTTTCTGACG ACGACGACGT CCAGCTTGTG GCCACTACCT  CAACGCGGTC CTCCAGCTCG CGCCATCTGA AAACCCACCG CCGGAGCAGA
      5010       5020       5030       5040       5050        5060       5070       5080       5090       5100        5110       5120       5130       5140       5150        5160       5170       5180       5190       5200
AGTCGCAGTA AGAATGTCCG AAGCAGTGAC AGCAGCCCTT CTAGTCGAGA  GTCCTCGCGC CGACGCCGCC AGAAGAGCAG CAGGCTTTCC AGGGAGCCAT  CATCGAACCC GCCACGTAAG TCCCACCATT CGTCAACGCA GCGAAAAAAG  ACCCCCAAAA AGAGGAGGCG CAGCAAAAGT AGAAGTCGCT CTAAAAGCAT
      5210       5220       5230       5240       5250        5260       5270       5280       5290       5300        5310       5320       5330       5340       5350        5360       5370       5380       5390       5400
CCGGCGGAGT AGGAGCATAA GCATATTGAG AAACAACCGG CGGAGCAGAT  CGCGTAGCCC CAGCTGTCGA AATGCAGAAC AGCGACGCCA GCAGGACCGA  CGGCGCACTC CCACTAAAAA GTCCCACAAG CGGCACAAGC GACGCAGGCG  CAGCAGCTCC CCCTGAGAAT CTTCCTAGAG CTCCGCAACC ACCTCCACAT
      5410       5420       5430       5440       5450        5460       5470       5480       5490       5500        5510       5520       5530       5540       5550        5560       5570       5580       5590       5600
TTCAAAGTTT TTTTGTAGGT TACAAGTTGT ATCAAATAAA CCCCCAACAT  ATTTTTGTAA GCAAGCGATG TGGTTCCCAC AAAGGAGATG GGACTACGTG  GAAATGAATC CCCCATTAGT GGTTCATCTT CTGCGAAGTC GAGTTGGGGA  TGCCCTACGC GCTGTTTTTA TATTACGTTA TGTATTTATG TATGCATGTA
      5610       5620       5630       5640       5650        5660       5670       5680       5690       5700        5710       5720       5730       5740       5750        5760       5770       5780       5790       5800
TGTACATATG TATAAGTTCA ACAAATTGTA TGTAAAATCG AAAATCTACA  TCTTGGACCA ACACTTCGAG CGAGGGAAGT CCAAAATGTA GGAGTTCGAT  GCTCTATTGG TGTTTATAAT TTTATGTATT TTGTGTAAAA GTAACCTATG  TATAAATCCG TGTAAGTAGC TTTACGATCT GGAAATTCCG TATTACGCCA
      5810       5820       5830       5840       5850        5860       5870       5880       5890       5900        5910       5920       5930       5940       5950        5960       5970       5980       5990       6000
GTGTTCAGAT TCTTCCTGCC CACTGCGCAC AGTGGGTATA AAAGCGCGGT  GGCAGTGAGC TCATCGATTC AGTTTAATTT GGATGCTCCG CACGACTCGA  CGATGCTTCA CTCAAACTAT CTTTTTCTGA TCGGCGGTTG CTGCCTGGTT  TTCGCTGTCT CAGCGCAGAT TCCCATCAGT AAGTGCATCC ATTCCTTGAA
      6010       6020       6030       6040       6050        6060       6070       6080       6090       6100        6110       6120       6130       6140       6150        6160       6170       6180       6190
GAAACCCGAT CCCTAGCCTC ACGCCCCTTT AACAAGCTTC TGACAATCCC  TAGCGCCGCG ACGTTGCCCG GCGAACGAGA CTTTCCTGGC CTGTGGTCCC  GATTGTCAAA CGGAGTGCGC CACGCTAGGA AAGCCCTGTC TGGTCAGGCA  CATCCGATGT CCAGATCCAT GCTACTGCAA CAAGGGCTTC G
```

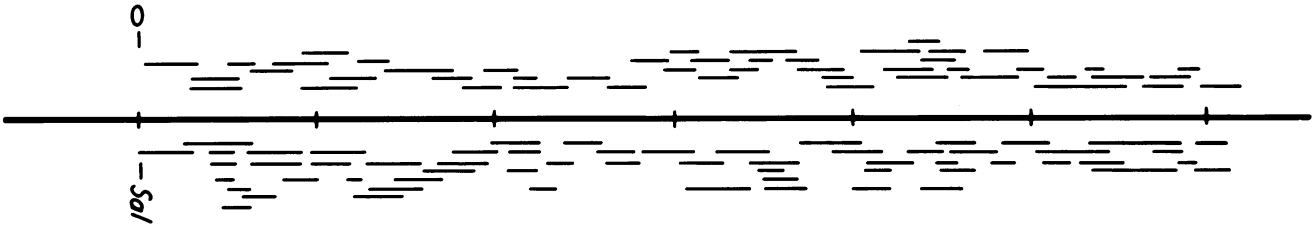

Fig. 1. DNA sequence of the 6191-base chromosomal segment containing the *suppressor-of-white-apricot* gene. The sequenced interval is numbered from the first base of the *Sal*I cleavage site (coordinate 1) through the last base of the *Nru*I cleavage site (coordinate 6191); these two cleavage sites bound the interval previously shown by gene transfer (Zachar *et al.*, 1978b) to contain $su(w^{a})$. This coordinate system will be used throughout this and the accompanying paper (Zachar *et al.*, 1987a). Sequence was determined for both strands and each restriction cleavage site internal to the sequenced interval was sequenced across. The sequence runs used are diagrammed as short, narrow lines. The bold line, calibrated in 1-kb intervals beginning at the *Sal*I site, represents the sequenced interval. The sequence shown is the wild-type allele parental to the $su(w^{a})^{hd7}$ P element insertion allele (Zachar *et al.*, 1987b) inferred from the sequence of the $su(w^{a})^{hd7}$ allele. The $su(w^{a})^{hd7}$ P element insertion is associated with a target site of duplication of bases 334–341 (TACTACAT). The ends of this P element are conventional (our unpublished results) and this mutation reverts at high frequency in PM dysgenic hybrids by P element excision (Zachar *et al.*, 1987b). The presumptive polyadenylation signal terminating a transcription unit immediately upstream of $su(w^{a})$ (the B transcription unit; Zachar *et al.*, 1987b) is at coordinate 138.

ing protection products from deletion protectors whose breakpoints occur within exon sequences. See Figures 3 and 5 for examples and diagrams. The presence of an exon identified by S_1 protection in a $su(w^{a})$ transcript is assessed by hybridization to Northern transfers with a probe segment subsumed by the exon sequences (Figure 4). This approach can be efficiently applied to analysis of a number of alternatively processed transcripts from a variety of developmental stages. Results of analysis of the 3′ region of the gene (in which all three transcripts are indistinguishable) are shown in Figure 3. Results of analysis of the 5′ region in which the three transcripts differ in structure are shown in Figure 5. Interpretation of most of the results of this analysis is straightforward and is recorded in Table I and diagrammed in Figure 6.

Conventions for naming of exons are as follows. All exons shared by more than one transcript class are numbered according to their occurrence in the 3.5-kb RNA. Exons in the 4.4- and 5.2-kb RNAs representing composites of exons from the 3.5-kb

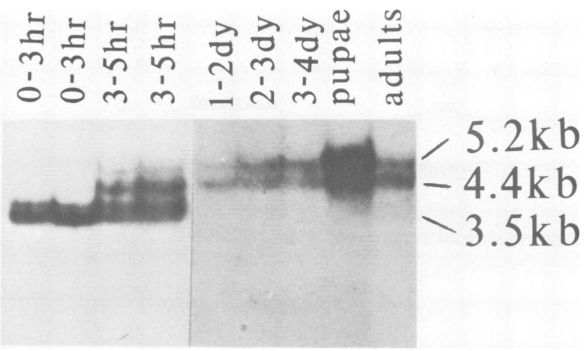

Fig. 2. Northern analysis of developmental pattern of polyadenylated *su(w^a)* transcripts. The probe extends from coordinates 2022 to 341 and is homologous to transcript orientation 1−6191, 5′ to 3′. Ten micrograms of polyadenylated RNA was loaded for the 3- to 5-h embryo samples and 2−3 μg for all other samples. Positions of the three major transcripts (3.5-, 4.4- and 5.2-kb RNAs) are indicated.

RNA are numbered to indicate this. Thus, the first exon of the 4.4-kb RNA is numbered 1−2 to indicate that it consists of exons 1 and 2 and intron 1 of the 3.5-kb RNA. The first exon of the 5.2-kb RNA is numbered 1−3 to indicate that it consists of exons 1, 2 and 3 and introns 1 and 2 of the 3.5-kb RNA (see Figures 3−5 for diagrams).

Several details of these results require additional comment.

(i) The 4.4- and 5.2-kb RNAs begin at the same site as the 3.5-kb RNA as evidenced by the following. Pupal RNA preparations (consisting of equal proportions of the 4.4- and 5.2-kb RNAs; Figure 2) show precisely the same S_1 protection pattern as precellular blastoderm embryonic RNA (consisting of the 3.5-kb RNA) with protectors spanning the 5′ ends of the RNAs (Figure 5, panel D; Table I). Probe fragments restricted to this 5′ terminal region (including protectors v and w in Figure 5) hybridize equivalently to all three transcript classes (Figure 4 and results not shown).

(ii) The common site of initiation of these three transcripts maps at or very near coordinate 333. This start site is 30 bases 3′ to a presumptive TATA sequence (TATAACT) beginning at coordinate 303. [In addition, an S_1 protection product suggesting a minor class of initiation events around coordinate 355 is seen from all RNA preparations (Figure 5, panel D, and results not shown).]

(iii) Exon 1−3 is sufficiently large that its structure must be reconstructed from a combination of measurements (Figure 5; Table I). First, exon 1−3 extends from its start site (coordinate 333) to the end of protectors p, q and t at coordinate 2022 as demonstrated by the larger protection product in channels pP, qP and tP in panel C of Figure 5. Second, exon 1−3 extends beyond 2022 to the end of protectors m and n at coordinate 2627 as demonstrated by the 1430-base protection product in channel mP and the 605-base protection product in channel nP in Figure 5, panel G. Third, exon 1−3 extends beyond coordinate 2627 to end at coordinate 2840 as demonstrated by the 445-base protection product in channels eP and oP in Figure 5, panel F.

(iv) Exon 1−3 is identified as a component of the 5.2-kb RNA based on hybridization to the 5.2-kb RNA, but not the 4.4-kb RNA, by a probe segment defined by coordinates 2395 and 2022 (Figure 4). This segment is homologous to exon 1−3 but not to exons 1−2, 2 or 3.

(v) Exon 1−2 is identified as a component of the 4.4-kb RNA by elimination. Probes contained within the interval of overlap between exons 1−2 and 1−3 hybridize to both the 4.4- and 5.2-kb RNAs (Figure 4). The results described in item (iv) above demonstrate that exon 1−3 is a component of the 5.2-kb RNA. Thus, exon 1−2 is identified as a component of the 4.4-kb RNA.

(vi) A minor protection product is seen in the pupal sample in channel pP (Figure 5, panel D) co-migrating with the exon 1 protection product from the 3.5-kb (embryonic) RNA. A similar minor protection product is reproducibly seen using larval RNA (results not shown). These results and those of Northern analysis (see Zachar *et al.*, 1987a,b) indicate that the 3.5-kb RNA is present as a minority species during much or all of postcellular blastoderm development.

(vii) Size and placement of exon 2 of the 3.5-kb RNA on the interval between coordinates 1680 and 1870 are based on the following observations in addition to those summarized in Table I (results not shown). A 180-base protection product is observed using the 3.5-kb (embryonic) RNA, protector fragments extending from coordinate 2022 to 1680 and from coordinates 2627 to 1198 and probe fragments extending from coordinate 1198 to 2627. Further, when a protector extending from coordinate 2022 to 1770 is used, the exon 2 protection product is reduced to a size below that detectable in our experiments (less than ~100 bases).

(viii) Exons 6a and 6b partially overlap. In addition to the results in Table I, the placement of exon 6b is based on elimination of its protection product in channel bP of panel E. Exons 6a and 6b are present in equal amounts in embryonic (3.5-kb RNA) and pupal (4.4- and 5.2-kb RNAs) preparations and are likely to represent alternative processing products distinguishing subclasses of the 3.5-kb RNA and one (and probably both) of the 4.4- or 5.2-kb RNAs. The representation of exons 6a and 6b in the various *su(w^a)* transcripts will be discussed in detail below.

(ix) Placement of exon 7 is refined to the interval between coordinates 4355 and 4560 by the observation that exons 7 and 8 must not overlap as they are represented in the same transcripts (Figure 4).

(x) Exon 8 ends at or near coordinate 5460. A conventional polyadenylation signal begins at coordinate 5435 (Figure 1).

(xi) The question arises as to the structure of the minor 3.7-kb species seen in precellular blastoderm samples (Figure 2). Each tested probe segment showing homology to the 3.5-kb RNA is also homologous to the 3.7-kb RNA (Figure 4 and results not shown). Moreover, we detect no extra S_1 protection products likely to account for the 3.7-kb species (Figures 3 and 5 and results not shown). Such additional protection products would be detectable in spite of the lower abundance of the 3.7-kb transcript if they were greater than ~300 bases in size. Thus, our results indicate that all of the larger exons of the 3.5-kb RNA (exons 1, 3, 4, 5 and 8) and probably the smaller exons are represented in the 3.7-kb RNA. While additional experimentation will be required to establish structure of the 3.7-kb RNA in detail, it is likely to be identical to that of the 3.5-kb RNA with the inclusion in the 3.7-kb RNA of an additional small exon(s) undetected by our S_1 protection studies. For simplicity, we will discuss *su(w^a)* expression henceforth in terms of the 3.5-kb RNA. However, the interpretation of our results would not be significantly altered by assuming that either the 3.7-kb and/or the 3.5-kb RNAs is necessary for *su(w^a)* function (see below).

Sequence analysis of cDNA clones

We sequenced two cDNA segments from a 0- to 3-h embryonic library (containing primarily the 3.5-kb RNA). One of these segments covers introns 1−3 and the other introns 4−7

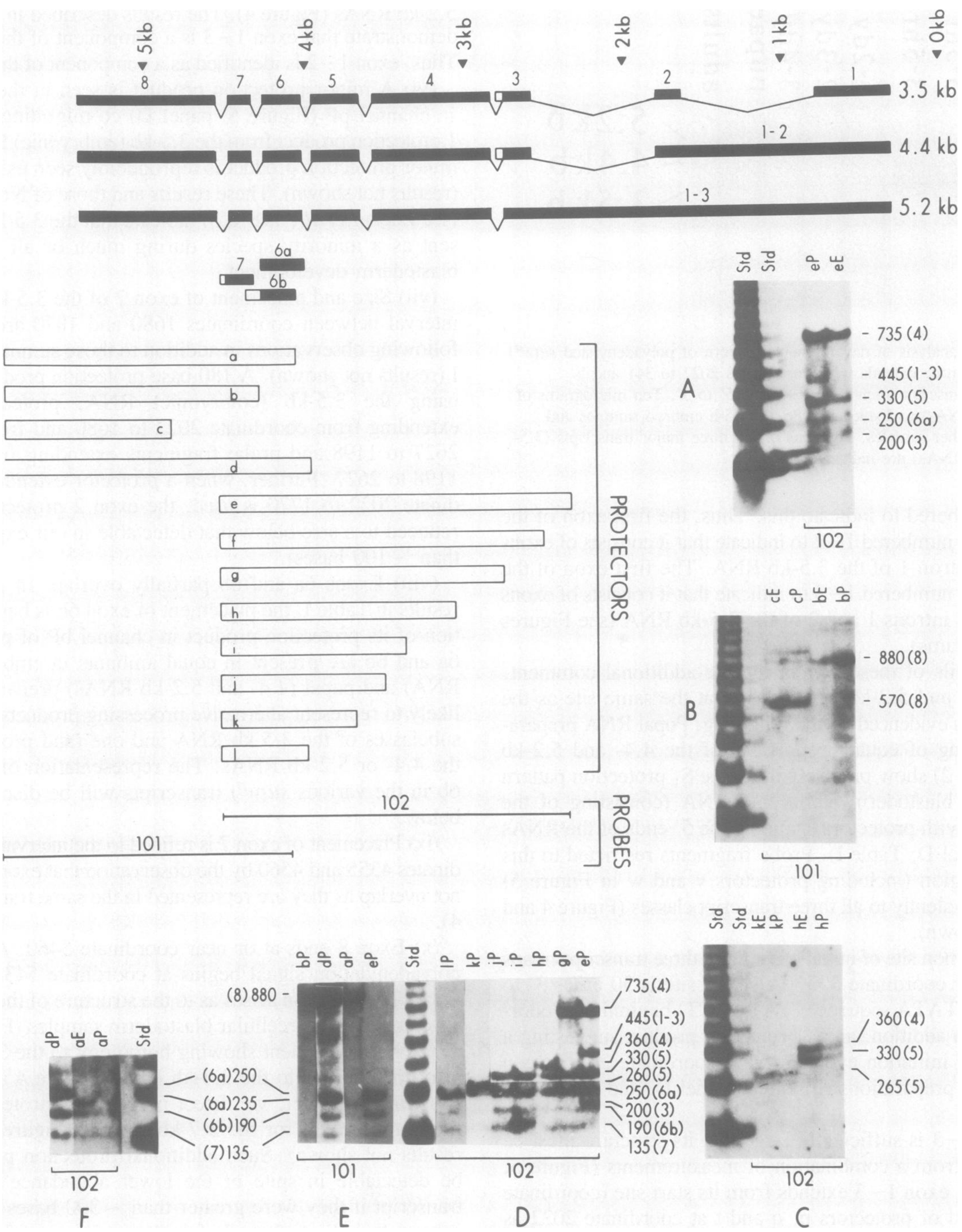

Fig. 3. S_1 protection analysis of the 3′ portions and *su(w^a)* transcripts. At top is a diagram of the structures of the three major *su(w^a)* transcripts projected on the standard coordinate system (Figure 1) and reconstructed from the S_1 protection results shown in Figures 3 and 5 (exons are solid bars and introns are diagonal lines; transcripts oriented 5′ to 3′, right to left). In most cases, exon boundaries are defined by S_1 protection with an uncertainty of ~15 bases. For exons 3, 6b and 7, where placement of these boundaries is less precise than 15 bases, open segments are appended to the exon to indicate uncertainty. The diagrammed transcript structures incorporate the precisely placed exon 6a rather than less precisely placed alternative exon 6b (see text). Also see Table II and Figure 6 for results of cDNA sequence refining the placement of the exon boundaries. S_1 protection was carried out as described in Materials and methods. Dimensions of the DNA segments used as protectors and as sequence probes are diagrammed. The precise dimensions of these segments are as follows (polarity 5′ to 3′, first coordinate to second coordinate): **Protectors—a** (5857–4095), **b** (5857–4353), **c** (5857–4887), **d** (6191–3998), **e** (4633–2395), **f** (4633–2643), **g** (4633–2846), **h** (4633–3261), **i** (4633–3493), **j** (4633–3538), **k** (4633–3738) and **l** (4088–4633). **Probes—101** (4095–5857) and **102** (2395–4633). The size of each major protection product in bases is indicated by appended number and is followed in parenthesis by the exon to which the protection product corresponds. RNAs (P for pupal and E for precellular blastoderm embryonic) and protecting DNA fragment (a–l) are indicated above each experimental channel. Probes used are indicated under each experimental filter. Molecular weight standards are a 123-base ladder (Materials and methods). Protection products corresponding to the small exons 6a, 6b and 7 are visible on longer autoradiographic exposure of **panel C**, channels bE and bP of **panel B** and channel kP in **panel D** (results not shown). Varying autoradiographic exposures of different portions of panels D–F are used to allow visualization of products of different autoradiographic intensity.

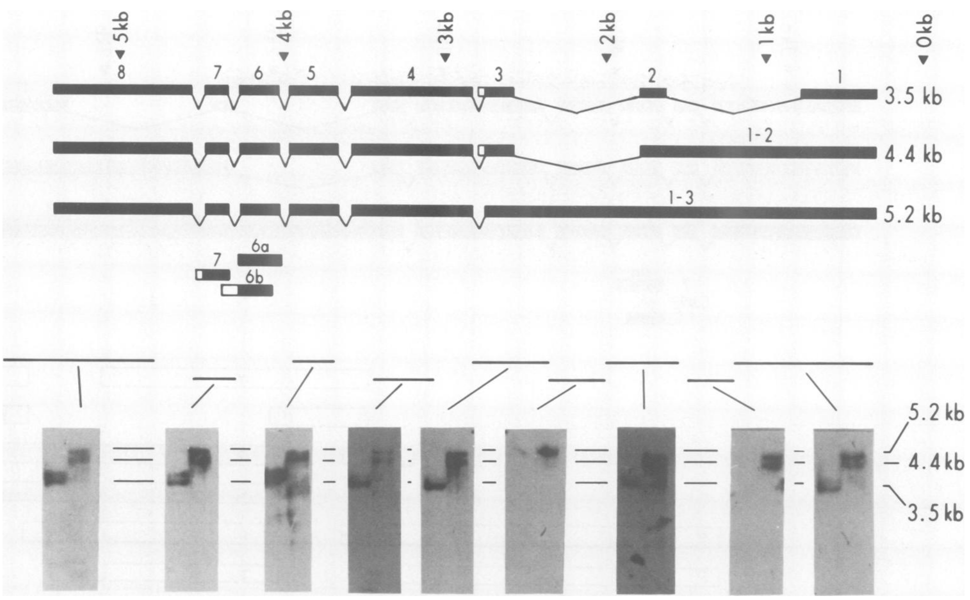

Fig. 4. Northern analysis of presence of exons in each of the major *su(w^a)* transcripts. Each Northern panel contains one channel (**left**) of 0- to 3-h embryonic RNA and one (**right**) of pupal RNA. Each filter was probed with the segment indicated by the connected horizontal line.

(Materials and methods). These studies allow precise placement of splice junctions and thus reconstruction of the sequence of the 3.5-kb transcript. This transcript contains a 2892-base open reading frame (ORF) extending from an AUG at coordinate 528 (~ 195 bases 3′ to the 5′ end of the RNA) to a UGA at coordinate 5364 (~ 95 bases 5′ to the 3′ end of the RNA). We interpret this structure to be that of a conventional mRNA. These results are recorded in Table II and diagrammed in Figure 6.

Our S_1 protection studies demonstrate that some combination of exons 6a, 6b and 7 are present in each *su(w^a)* transcript, but do not establish which of the possible arrangements actually exists. The results of cDNA sequence allow resolution of this issue as follows. Exons 6a and 6b overlap and are presumably in different classes of mature transcripts. These data demonstrate the existence of exon arrangment 6a−7 (Table II; Figure 6). The only arrangement containing exon 6b that would nearly comigrate with transcripts containing the 6a−7 arrangement is 6b−7. (The 6a−7 and 6b−7 arrangements would differ in size by ~65 bases and this difference would not result in resolution in the 3.5- to 5.2-kb mol. wt range under our conditions.) The alternative splice(s) producing exons 6a and 6b is not subject to the regulation affecting the first two introns as demonstrated by their similar stoichiometries in precellular blastoderm and pupal RNAs (Figure 3; results above). For simplicity, we will not distinguish between the 6a and 6b transcript classes in our discussions below.

Our S_1 protection studies demonstrate that the three major transcript classes are indistinguishable in structure in their 3′ portions at a very high level of resolution (results above; Figures 3 and 5). Thus, the sequence of cDNA clones of the 3.5-kb RNA can be used to reconstruct the sequence of the corresponding portions of the 4.4- and 5.2-kb RNAs (Table II; Figure 6).

Structure of the presumptive su(w^a) protein products

The conceptual translation product of the long ORF extending through most of the length of the 3.5-kb RNA is shown in panel A of Figure 7. This peptide sequence begins at the methionine corresponding to the first in-frame AUG at coordinate 528 (Figure 6). The conceptual translation product of the shorter ORF beginning at this same AUG in the 4.4- and 5.2-kb RNAs (Figure 6) is shown in panel B of Figure 7.

Discussion

Developmental regulation of su(w^a) expression

Our results define the complete sequence of the *su(w^a)* locus including the primary transcription unit. Our cDNA sequence studies define the seven introns whose removal produces the 3.5-kb precellular blastoderm *su(w^a)* mature RNA species. S_1 protection experiments demonstrate that the 4.4- and 5.2-kb postcellular blastoderm mature RNAs are derived from the same primary transcript as the 3.5-kb RNA and differ from it by retaining the first or the first and second introns respectively (Figure 6). Thus, the dramatic change in *su(w^a)* transcript pattern occurring during early embryogenesis (Figure 2) results from specific, developmentally programmed cessation of splicing of the first two introns. Previous studies indicate that *su(w^a)* is transcribed in most or all cells in the organism (Zachar *et al.*, 1987b). Thus, these developmental changes in splicing are likely to be temporal changes involving the same lineages rather than tissue-specific differences in RNA processing.

Functional significance of developmental regulation of su(w^a) splicing

Three rationales for the unusual mode of post-transcriptional regulation of *su(w^a)* transcript pattern are consistent with our results.

(i) The various mature *su(w^a)* transcript classes could code for different polypeptides (Figures 6 and 7). The structure of the 3.5-kb RNA is consistent with its being a conventional mRNA whose translation is initiated at the AUG at coordinate 528 (~ 195 bases into the RNA). In contrast, the two postcellular blastoderm mature RNAs (4.4 and 5.2 kb) have highly unusual structures for presumptive mRNAs (Figure 6). In light of the 3.5-kb RNA, it is economical to suppose that the functional peptide coding sequence of the 4.4- and 5.2-kb RNA is the relatively short open reading frame whose translation is initiated at the AUG at coordinate 528. However, each of the 4.4- and 5.2-kb RNAs con-

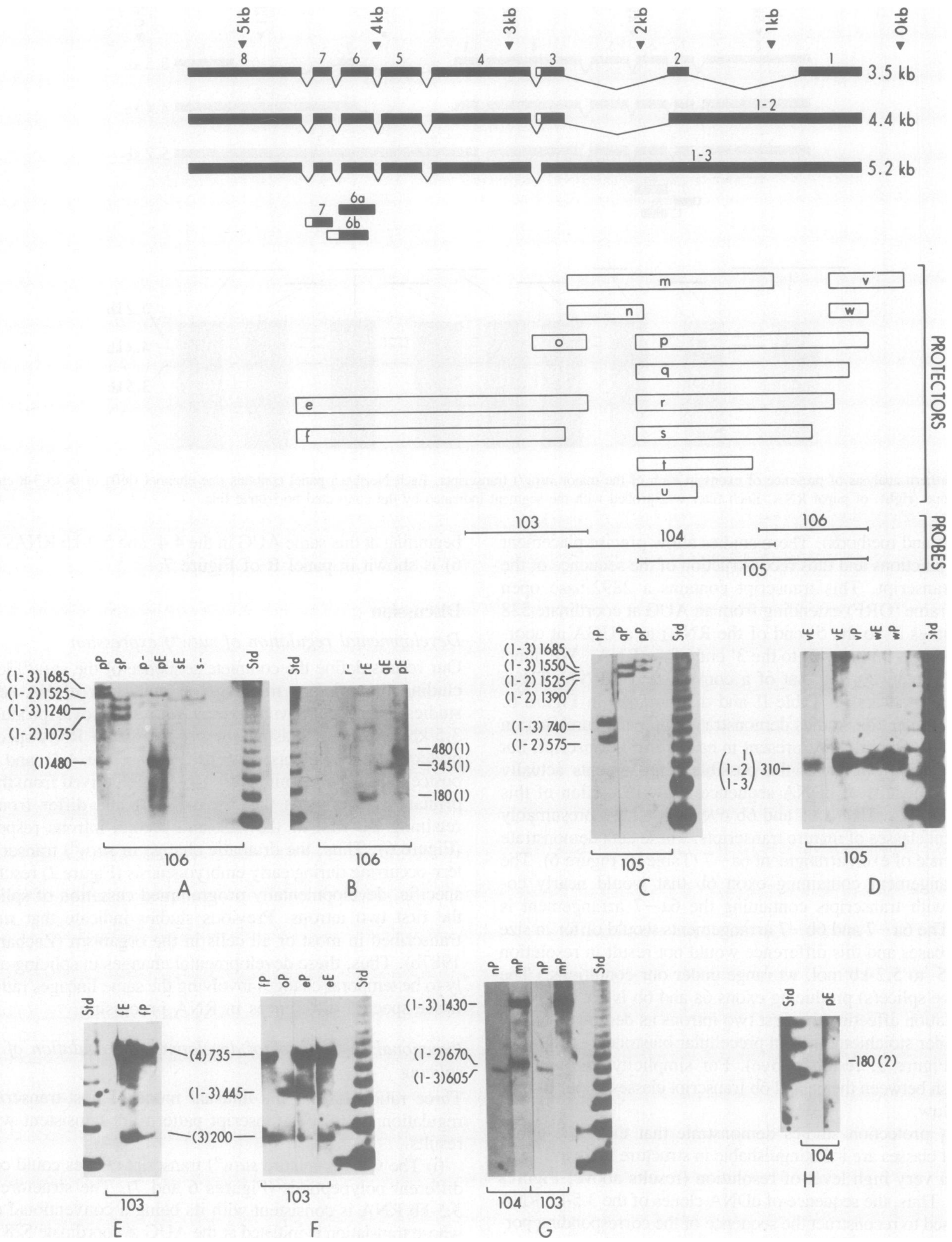

Fig. 5. S_1 protection analysis of the 5′ portions of *su(w^a)* transcripts. Structures of the three major *su(w^a)* transcripts are diagrammed (5′ to 3′, right to left) as in Figure 3. S_1 protection was carried out as described in Materials and methods. Dimensions of the DNA segments used as protectors and as sequence probes are diagrammed. The precise dimensions of these segments are as follows (polarity 5′ to 3′, first coordinate to second coordinate): **Protectors—e** (4633–2395), **f** (4633–2643), **m** (2627–1198), **n** (2627–2022), **o** (2868–2395), **p** (2022–341), **q** (2022–476), **r** (2022–629), **s** (2022–788), **t** (2022–1288), **u** (2022–1678), **v** (637–1) and **w** (637–341). **Probes—103** (2395–3312), **104** (1198–2627), **105** (341–2022) and **106** (341–1029). The size of each major protection product in bases is indicated and appended in parentheses is the exon to which the protection product corresponds. RNAs (P for pupal and E for precellular blastoderm embryonic) and protecting DNA fragment (e–w) are indicated above each experimental channel. Probes used are indicated under each experimental filter. Molecular weight standards are a 123-base ladder (Materials and methods).

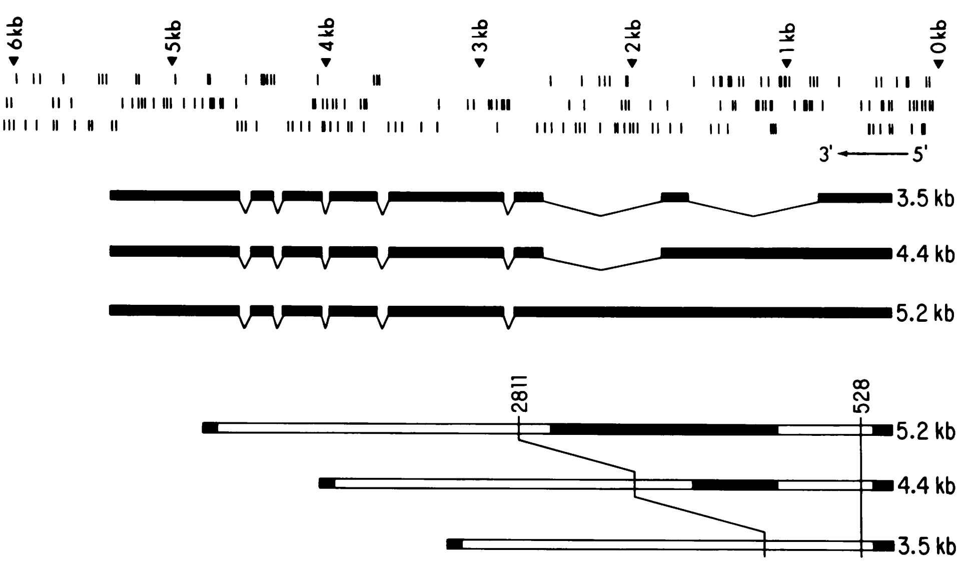

Fig. 6. Structures of the three major *su(w^a)* transcripts based on S_1 protection (Table I) and cDNA sequence (Table II) results. Shown at top is a conventional ORF diagram of the *su(w^a)* sequence (Figure 1). The three reading frames (5' to 3', right to left; defined beginning at base 1 of the sequenced interval) are indicated as three lines of symbols (top line is frame 1, middle is frame 2 and bottom is frame 3). A vertical line is placed at each translation termination codon in the reading frame. Thus, extended segments of uninterrupted potential peptide coding sequence are indicated by intervals free of vertical lines. Immediately below the ORF diagram are the structures of the three mature *su(w^a)* transcripts diagrammed as in Figure 3. At bottom are the structures of the three mature *su(w^a)* transcripts diagrammed with introns removed. These are drawn to the same scale as the upper diagrams and are aligned at their 5' ends. The open portions of the bars indicate ORFs longer than 450 bases. The positions of the first in-frame AUG in each ORF (AUGs at coordinates 528 and 2811, respectively) are indicated. The 3.5-kb RNA contains a single ORF extending from an AUG at coordinate 528 through a UGA at coordinate 5364 (also see Table II). This extended ORF remains separated in two or three non-continuous components when one or both of the first two splices fails to occur resulting in the 4.4- and 5.2-kb mature RNAs.

```
A   1  MLPYNVRNAG GGSVGGILRR TGQGSGTGST ILGNGNSPGA LGAGKVSSSS LENHRQPPLE
   61  LLVFGYACKI FRDDEKAREM DHGKQLIPWM GDVNLKIDRY DVRGALCELA PHEAPPGGYG
  121  NRLEYLSAEE QRAEQLCEEE RYLFLYNNEE ELRLRQEEDL KRLQQETSGG CFSQVGFQYD
  181  GQSAASTSIG GSSTATSQLS PNSEESELPF VLPYTLMMAP PLDMQLPETM KQHAIIEKTA
  241  RFIATQGAQM EILIKAKQAN NTQFDFLTQG GHLQPYYRHL LAAIKAAKFP PAPQTPLDQQ
  301  NTDKEAPSAD DHSEEVAGGR RNPNQVVITV PTIKYKPSAN CAYTQLISKI KGVPLQAVLQ
  361  EDESSNPGNS QHSGGTASPA LSCRSEGHNS QGGEFTPVLL QYNGSTFTHE EESSNREQQD
  421  DNDVNGGEPP QVELLKNTSA LALAQNYSSE SEEEEDQVQP EKEEEKKPEP VLTFPVPKDS
  481  LRHIIDKTAT YVIKNGRQFE ETLRTKSVDR FSFLLPANEY YPYYLYKVTG DVDAASKEEK
  541  TRKAAAVAAA LMSKKGLSFG GAAAAVSGSN LDKAPVSFSI RARDDQCPLQ HTLPQEASDE
  601  ETSSNAAGVE HVRPGMPDSV QRAIKQVETQ LLARTAGQKG NITASPSCSS PQKEQRQAEE
  661  RVKDKLAQIA REKLNGMISR EKQLQLERKR KALAFLNQIK GEGAIVGSAV PVVGPNPPES
  721  AAGAATADSG DESGDSVRSI PITYFGPDDD DEVGEQRPEM RLIGSTQKDE EDDDEEDGGD
  781  LEKYNLLNDD STNTFTSKPV LPPTAAPPPA AVLLSDDDDV QLVATTSTRS SSSRHLKTHR
  841  RSRSRSKNVR SSDSSPSSRE SSRRRRQKSS RLSREPSSNP PRKSHHSSTQ RKKTPKKRRR
  901  SKSRSRSKSI RRSRSISILR NNRRSRSRSP SCRNAEQRRQ QDRRRTPTKK SHKRHKRRRR
  961  SSSP
```

```
B   1  MLPYNVRNAG GGSVGGILRR TGQGSGTGST ILGNGNSPGA LGAGKVSSSS LENHRQPPLE
   61  LLVFGYACKI FRDDEKAREM DHGKQLIPWM GDVNLKIDRL IIGTNPSNNY TINQYAHHCH
  121  KHTHTATYPP MSCPVRTTLP HTAAVKLLAE SDTYIYVGIV IDGIVIMALE DRACRYL
```

Fig. 7. Conceptual translation products of *su(w^a)* RNAs. **Panel A**: amino acid sequence of the translation product of 3.5-kb RNA (Figure 6; Table II) is given using the conventional single-letter amino acid code. Translation is initiated at the first in-frame AUG (coordinate 528). Translation initiated at the AUG at coordinate 2811 (see Figure 6) would produce the peptide initiated at the methionine at amino acid 217 in the sequence shown. **Panel B**: amino acid sequence resulting from translation of the 4.4- and 5.2-kb RNA initiated at the AUG at coordinate 528.

tains a long open reading frame in its 3' portion that, in the absence of information about the 3.5-kb RNA, would be presumed to be the protein coding region (Figure 6). On this latter assumption, the 4.4- and 5.2-kb RNAs would be messages with unusually long 5' untranslated regions containing long secondary open reading frames. (The first in-frame AUG in the 3' ORF in each of these RNAs is at coordinate 2811 corresponding to a presumptive 5' untranslated region of 1710 bases for the 4.4-kb RNA and 2479 bases for the 5.2-kb RNA.)

(ii) All three major transcript classes might code for the same polypeptide. The simplest version of this hypothesis assumes that translation initiates at the AUG at coordinate 2811 and extends through the common 3' portion of the three RNAs (Figure 6). On this hypothesis the alternative splicing distinguishing *su(w^a)* transcripts would presumably serve to create mature RNAs whose translation behavior differs as a result of their very different 5' untranslated regions. The long, in-frame segment of peptide coding sequence (216 codons) 5' to the AUG at 2811 in the case of the 3.5-kb RNA argues against this hypothesis.

(iii) As the 3.5-kb RNA has a structure consistent with its being a conventional message, it is conceivable that the 4.4- and 5.2-kb RNAs represent by-products of post-transcriptional repression of expression of the 3.5-kb RNA. On this model the 4.4- and 5.2-kb RNAs need not have any functional peptide coding capabilities. It is noteworthy in this regard that trace amounts of the 3.5-kb RNA persist throughout postcellular blastoderm development when the 4.4- and 5.2-kb RNA are the preponderant species (Figure 5, panel A). Results reported in the accompanying paper (Zachar *et al.*, 1987a) demonstrate that this last hypothesis is likely to be correct.

Detailed mechanism of developmental regulation of su(w^a) splicing

Two general classes of mechanisms for regulation of splicing of the first two *su(w^a)* introns are consistent with our results. First, the 'activation hypothesis' supposes that a factor(s) necessary for these splices is present at high levels only transiently during precellular blastoderm development. Second, the 'repression hypothesis' supposes that these splices are potentially efficient throughout development but a factor(s) repressing them is present after cellular blastoderm. Our results do not distinguish between these hypotheses. However, the results reported in the accompanying paper indicate that the repression hypothesis is correct.

Table I. Structural analysis of *su(w^a)* transcripts by S_1 protection

Exon	Protection products[a]	Exon size/placement[b]
1	480 (pE, A, 5)	480 (333–813)
	480 (pE, B, 5)	
	345 (qE, B, 5)	
	180 (rE, B, 5)	
	310 (vE and wE, D, 5)	
2	180 (pE, H, 5)	180 (1680–1870)[c]
1–2	1525 (pP, C, 5)	1530 (333–1865)
	1390 (qP, C, 5)	
	575 (tP, C, 5)	
3	200 (all channels, F, 5)	200 (2643–2868)
	200 (fE and fP, F, 5)	
1–3	1685 (pP, C, 5)	2508 (333–2840)
	1550 (qP, C, 5)	
	740 (tP, C, 5)	
	1430 (mP, G, 5)	
	605 (nP, G, 5)	
	445 (eP and oP, F, 5)	
4	735 (eP and eE, A, 3)	735 (2885–3620)
	735 (eP and gP, D, 3)	
	360 (hP, D, 3)	
	360 (hP and hE, C, 3)	
5	330 (eP and eE, A, 3)	330 (3675–4005)
	330 (hE and hP, C, 3)	
	330 (eP–jP, D, 3)	
	265 (kE and kP, C, 3)	
6a	250 (all channels, D, 3)	250 (4080–4330)
	250 (eP, E, 3)	
	250 (dP, F, 3)	
	235 (aE and aP, F, 3)	
	235 (aP, E, 3)	
6b	190 (lP, D, 3)	190 (4095– ~4400)
	190 (dP, E, 3)	
	190 (dP, F, 3)	
	190 (aE and aP, F, 3)	
7	135 (all channels, D, 3)	135 (4355–4633)
	135 (bP, E, 3)	
	135 (aP and aE, F, 3)	
8	880 (bE and bP, B, 3)	880 (4560–5460)
	570 (cE and cP, B, 3)	

'Summarized are details of S_1 protection experiments shown in Figures 3 and 5.

[a] Protection products are given as follows: size in bases (channel, panel, figure). Thus, a 500-base protection product in channel pE in panel C of Figure 3 is indicated as follows: 500 (pE, C, 3).

[b] Exon dimensions are given as follows: size of exon (coordinates of exon). Thus, a 500-base exon mapping between coordinates 5025 and 5525 would be indicated as follows: 500 (5025–5525). Where the size of the exon is the same as the distance between the coordinates, the uncertainty in placement of the endpoints of the exon is ~15 bases. Where the interval between the coordinates is larger than the exon size the exon can map anywhere within the coordinate interval given. See text for further discussion of exon placement.

[c] Also see text for discussion of additional observations placing this exon.

It is noteworthy that splicing of the second intron continues to occur with some efficiency in postcellular blastoderm individuals, whereas removal of the first intron is very substantially depressed. While the different behavior of the first and second introns could be selectively significant in some way not currently apparent, the behavior of the second intron is equivalently interpretable as adventitious. The secondary and tertiary structure of an RNA molecule might influence efficiency of its splicing (see Padgett *et al.*, 1986, for a recent review). Removal of the first *su(w^a)* intron might substantially alter the folding of the second. We speculate that splicing of the first intron is the active target for *su(w^a)* regulation and removal of the second intron is inadvertantly affected.

The sequence of the first intron has an unusual feature. Introns 2–7 each contain a match to the *Drosophila* branch site consensus in the expected sequence interval (distances of 17–50 bases separating the 3′ end of the intron and the A at the presumptive branch site; see Padgett *et al.*, 1986, and references therein). In contrast, the nearest first intron match to the consensus occurs at a distance of 67 nucleotides (Table II). Thus, the first intron either uses an unconventional branch site sequence or uses a conventional sequence at an unconventional distance from the 3′ splice site. In either case, splicing of this intron would probably be slower (see Padgett *et al.*, 1986, for a review) and/or unusually sensitive to folding of the intron. This absence of a conventional presumptive branch site sequence is uncommon (for example, see Keller and Noon, 1984) and thus seems likely to be related to regulation of first intron splicing.

The nature of the su(w^a) protein product

The conceptual translation product of the presumptive *su(w^a)* mRNA (the 3.5-kb RNA; Figure 7) has an unusual structure. This protein has two recognizably distinct domains. The amino-terminal 85% (approximately amino acids 1–824) is modestly acidic (with a calculated pI of 5.5) while the carboxy-terminal 15% (approximately amino acids 825–964) is strikingly basic (29% arginine and 10% lysine; a calculated pI of 13.7). Further, this unusually basic carboxy-terminal domain is rich in hydroxylated amino acids (31% serine and 6% threonine). Computer-assisted screens of currently available protein sequence banks fail to detect proteins with extensive similarity to this conceptual translation product (our unpublished observations); however, limited similarity is detected between the unusual carboxy-terminal domain of this conceptual translation product and the very heavily phosphorylated storage protein, phosvitin (Byrne *et al.*, 1984). Perhaps this unusual domain in the presumptive *su(w^a)* protein is phosphorylated.

Experiments described in the accompanying paper (Zachar *et al.*, 1987a) indicate that the *su(w^a)* protein is a regulator of splicing (and, possibly, of other RNA processing events). It will be of interest to determine if the unusual structure of the *su(w^a)* protein is related to this regulatory property.

Materials and methods

DNA sequence determination

DNA sequence was determined by chain termination methods (Sanger *et al.*, 1977; Barnes *et al.*, 1983). Most of the sequence was determined by subdividing the 6191-base *su(w^a)* segment into several partially overlapping fragments and generating a nested deletion set for each of these fragments essentially as described by Barnes *et al.* (1983). Sets of deletion derivatives were chosen which removed progressive increments of 150–200 bases from the parental clone, beginning at the end juxtaposed to the sequencing primer binding site in conventional M13 cloning vectors. Sequencing of such a deletion set allows convenient reconstruction of the entire original fragment.

Northern and Southern analysis

Formaldehyde–agarose gels were used for Northern and Southern analyses (Maniatis *et al.*, 1982). Nitrocellulose transfers were probed with single-stranded M13 probes made according to Hu and Messing (1982) and used according to Bingham and Zachar (1985).

Molecular weight standards for RNA measurements were the following *Drosophila* transcripts (visualized by reprobing Northern filters with the appropriate sequences): *gypsy* (Marlor *et al.*, 1986), *copia* (Emori *et al.*, 1985), *white* (Pirrotta and Brockl, 1984; Levis *et al.*, 1984; Davison *et al.*, 1985). *Adh* (Benyajati *et al.*, 1983) and *rp49* (O'Connell and Rosbash, 1984). We estimate that our size measurements for RNAs are accurate to within less than ~10%.

Table II. Precise placement of *su(w^a)* introns from cDNA sequence

Exon	Transcript class (kb)	Exon size/placement[a]	Open reading frame[b]	Presumptive branch site[c]	3′ splice site[d]	5′ splice site[d]
1	3.5	490 (333–823)	3 (453–1058)	–	–	CAGgttaat
2	3.5	173 (1680–1853)	1 (1588–2019)	tttat (67)	tgaccacagATA	AAGgtgagc
1–2	4.4	1531 (333–1853)	3 (453–1058)	–	–	AAGgtgagc
3	3.5, 4.4	208 (2633–2841)	1 (2542–3648)	ccaat (17)	atcctgcagAGG	CTGgtgagt
1–3	5.2	2508 (333–2841)	3 (453–1058)	–	–	CTGgtgagt
4	3.5, 4.4, 5.2	720 (2911–3631)	1 (2542–3648)	ccaaa (38)	gctttccagCCC	AAGgtgggc
5	3.5, 4.4, 5.2	321 (3688–4009)	1 (3688–4047)	tgtat (25)	ttattttagAAA	AGGgtgagt
6a	3.5, 4.4, 5.2	251 (4070–4321)	1 (4050–4329)	ttaat (20)	tttttgtagCTC	AAGgtacga
7	3.5, 4.4, 5.2	129 (4388–4517)	2 (4079–4573)	ttaat (25)	gtcgtgaagCTG	AAGgtgagc
8	3.5, 4.4, 5.2	888 (4572–5460)	3 (4554–5363)	tgtat (23)	ttctaccagGTG	–

Branch site consensus: CTAAT
TCG C
g t a

Summarized are placements of splices based on sequence of cDNA clones (Materials and methods). cDNAs sequenced are clones of the 3.5-kb RNA. Positions of splices producing the 4.4- and 5.2-kb RNAs are inferred from the cDNA sequence based on S_1 protection results summarized in Table I. These results are diagrammed in Figure 6.

[a] Exon dimensions are given as in Table I as follows: size of exon (coordinates of exon).

[b] Reading frame (1, 2 or 3) refers to frame numbering from the first base of the sequenced interval. This convention is used in Figure 6. The dimensions of the ORF in the chromosomal DNA sequence subsuming the indicated exon is given in parentheses following the reading frame designation. Determination of the relationship between reading frames in exons juxtaposed by splicing can be made by recalling that a base whose coordinate value divided by three yields a value of an integer and a 1/3 is the first base of a triplet in reading frame 1, an integer and 2/3 the first base of a triplet in reading frame 2 and a precise integer the first base of a triplet in reading frame 3. Exon 1 contains a presumptive initiator AUG at coordinate 528 and exon 8 contains a presumptive translation termination signal at coordinate 5364.

[c] Shown is the first sequence more than 12 bases 5′ to the corresponding 3′ splice site that matches the *Drosophila* branch site consensus given at the bottom. Distances from the presumptive branch acceptor A through the last base of the intron, inclusive, are given in parentheses following the branch site sequence. The consensus given in bold, capital letters consists of the majority bases at each position, the sequence in capital letters immediately below the bold are frequent minority choices at the corresponding positions and the sequence in lower case letters are rare minority choices at these positions (Keller and Noon, 1984).

[d] 3′ splice site indicates cleavage at the 3′ end of the intron preceding the exon in question and 5′ splice the cleavage at the 5′ end of the intron following the exon. Capital letters indicate exon sequences and lower cases letters intron sequences.

Molecular weight standards for Southern analysis were a commercially available (BRL) 123-base ladder. These fragments were end labeled by conventional Klenow filling which adds two bases to each fragment class. Thus, the size of the *n*-mer in this ladder is (*n*) (123)+2. Size estimates using this standard ladder are reliable within less than ~15 bases in the mol. wt range of 125–900 bases.

S_1 protection

For panel A in Figure 5, S_1 protection was done according to the procedure of Bingham and Zachar (1985). This procedure occasionally introduces artifactual bands resulting from uncharacterized contamination of the M13-cloned segments used as protectors. (These bands are present in mock reactions in which all components of the reaction other than RNA are included. For example, see the RNA-independent, low stoichiometry bands at high mol. wt in panel A of Figure 5.)

The following modified procedure for S_1 protection eliminates these artifactual species and has been used in all cases here other than panel A of Figure 5: ~100 ng of an M13-cloned segment in 2 μl of TE is mixed with 2.5 μl of 4.4% SDS in a 0.5-ml Eppendorf tube and heated for 10 min at 65°C. The tube is spun briefly to recover condensate. Simultaneously, 10–30 μg of polyadenylated RNA is collected from an ethanol precipitate by centrifugation in an unsiliconized glass tube and the pellet is washed by gently filling the tube once with 95% ethanol at −20°C. The RNA pellet is thoroughly drained, dried briefly *in vacuo* and resuspended in 10 μl of 75% formamide buffered at pH 7.0 with 20 mM MOPS, 5 mM sodium acetate, 0.1 mM EDTA. The 10-μl RNA sample is added to the SDS–DNA mixture (above) together with 2 μl of 5 M NaCl. This hybridization reaction is incubated at 37°C for 12–24 h. After hybridization the reaction is diluted to 1.5 ml with 0.5 M NaCl, 0.5% SDS, 10 mM Tris, 1 mM EDTA pH 7.4 (high salt buffer) and run over an oligo(dT)–cellulose column (~0.2 ml bed volume). The column is washed with an additional 0.3 ml of high salt buffer to elute components not bound to the column by virtue of hybridization to bound polyadenylated RNA. The bound material is then eluted in 0.5–0.8 ml of 1 mM Tris, 0.1 mM EDTA (low salt buffer). The eluted material is ethanol precipitated and resuspended in 50 μl of low salt buffer. This resuspended material is mixed with one volume of 2 × S_1 buffer and digested as described previously (Bingham and Zachar, 1985). S_1-resistant material is recovered by ethanol precipitation and analyzed by probing a Southern transfer of the material with selected sequence probes (see figure legends).

All solutions to which the RNA is exposed in this procedure are treated by making 0.1% in diethylpyrocarbonate (Sigma), incubating overnight and autoclaving with the exceptions of the M13 DNA and formamide solutions. Formamide (Fluka) is deionized before use by rolling for 20 min at room temperature with ~0.1 volume of a mixed bed resin (BioRad), aliquoting and storing frozen at −80°C; residues of aliquots are discarded, not refrozen.

Sizes of S_1 protection products can generally be measured with an uncertainty of ~15 bases under our conditions. Breakpoints of the deletion variants used are known precisely by DNA sequence analysis. Thus, locations of exon boundaries can generally be determined with an uncertainty of ~15 bases. Moreover, two protection products co-migrating in the mol. wt range of ~100–500 bases under our conditions are identical in mol. wt with an uncertainty of ~5 bases.

Retrieval of cDNA segments

cDNA segments were retrieved from a λ-gt10 library prepared from 0- to 3-h embryo RNA. This library contains primarily fragments of less than full length and we retrieved the largest segment covering the first three introns (0.9 kb) and the largest segment covering the last four introns (2.6 kb). The first of these was sequenced through the interval corresponding to coordinates 700–3040 and the second through the interval corresponding to coordinates 3470–4620. This allows precise placement of all seven introns identified by S_1 protection (Table II). The cDNA sequence differs from our chromosomal sequence by G to A change at coordinate 4302 and a T to C change at coordinate 3747. Neither difference changes the amino aicd sequence encoded by the long ORF in the 3.5-kb RNA (Figure 7) and both probably represent differences between *su(w^a)* alleles.

Acknowledgements

We are grateful to Carolyn Chapman, Inka Mims and Michael Belanich for helpful criticisms of a draft of this manuscript. We are grateful to Stephen Poole and Thomas Kornberg for making their embryonic cDNA library available to us. This work was supported by NIH grant GM320003 to P.M.B.

References

Barnes,W.M., Bevan,M. and Son,P.H. (1983) *Methods Enzymol.*, **110**, 343–380.

Bender,W., Akam,M., Karch,F., Beachy,P.A., Peifer,M., Spierer,P., Lewis,E.B. and Hogness,D.S. (1983) *Science*, **221**, 23–29.

Benyajati,C., Spoerel,N., Haymerle,H. and Ashburner,M. (1983) *Cell*, **33**, 125–133.

Bingham,P.M. and Zachar,Z. (1985) *Cell*, **40**, 819–825.

Byrne,B.M., von het Schip,A.D., van de Klundert,J.A.M., Arnberg,A.C., Gruber,M. and Geert,A.B. (1984) *Biochemistry*, **23**, 4275–4279.

Davison,D., Chapman,C., Wedeen,C. and Bingham,P.M. (1985) *Genetics*, **110**, 479–494.

Emori,Y., Shiba,T., Kanaya,S., Inouye,S., Yuki,S. and Saigo,D. (1985) *Nature*, **315**, 773–776.

Hu,N. and Messing,J. (1982) *Gene*, **17**, 271–277.

Keller,E.B. and Noon,W.A. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 7417–7420.

Levis,R., O'Hare,K. and Rubin,G.M. (1984) *Gene*, **38**, 471–480.

Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory Press, New York, pp. 202–203.

Marlor,R.L., Parkhurst,S.M. and Corces,V.G. (1986) *Mol. Cell Biol.*, **6**, 1129–1134.

Modolell,J., Bender,W. and Meselson,M.S. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1678–1682.

O'Connell,P. and Rosbash,M. (1984) *Nucleic Acids Res.*, **12**, 5494–5513.

Padgett,R.A., Grabowski,P.J., Konarska,M.M., Seiler,S. and Sharp,P.A. (1986) *Annu. Rev. Biochem.*, **55**, 1119–1150.

Parkhurst,S.M. and Corces,V.G. (1986) *BioEssays*, **5**, 52–57.

Parkhurst,S.M. and Corces,V.G. (1987) *EMBO J.*, **6**, 419–424.

Pirrotta,V. and Brocki,Ch. (1984) *EMBO J.*, **3**, 563–568.

Sanger,R., Nicklen,,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.

Varmus,H.E. (1983) In Shapiro,J. (ed.), *Mobile Genetic Elements*. Academic Press, New York, pp. 411–505.

Winston,F., Chaleff,D.T., Valent,B. and Fink,G.R. (1984) *Genetics*, **107**, 179–197.

Zachar,Z. and Bingham,P.M. (1982) *Cell*, **30**, 529–541.

Zachar,Z., Davison,D., Garza,D. and Bingham,P.M. (1985) *Genetics*, **111**, 495–515.

Zachar,Z., Chou,T.-B. and Bingham,P.M. (1987a) *EMBO J.*, **6**, 4105–4111.

Zachar,Z., Garza,D., Chou,T.-B., Goland,J. and Bingham,P.M. (1987b) *Mol. Cell Biol.*, **7**, 2498–2505.

Received on September 1, 1987; revised on October 12, 1987