The EMBO Journal vol.6 no.6 pp.1775–1779, 1987

The structure of the *Ultrabithorax* promoter of *Drosophila melanogaster*

Gena Saari and Mariann Bienz

Zoological Institute, University of Zürich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland

Communicated by H.Pelham

The sequence of 4118 nucleotides upstream of the putative initiator codon of the *D. melanogaster Ultrabithorax* protein was determined. The transcription initiation site for the corresponding mRNA was identified by S1 nuclease mapping and primed extension. It appears that all embryonic RNA products that encode the first protein exon are initiated approximately one kilobase upstream of the initiator methionine, suggesting a unique *Ultrabithorax* promoter. The unusually long mRNA leader is unspliced. Upstream of the initiator codon occur two additional methonine codons; the first one is followed by an open reading frame encoding a putative polypeptide of 69 amino acids. We discuss the role of the leader and 5′ flanking sequences with respect to transcriptional and posttranscriptional control of *Ultrabithorax* gene expression.

Key words: homeotic gene/*Ultrabithorax*/promoter sequence/leader peptide

Introduction

The bithorax complex (*BX-C*) in *Drosophila* is of crucial importance for normal development of thoracic and abdominal segments (Lewis, 1963, 1978; reviewed by Lawrence and Morata, 1983). It consists of three major genes (*Ultrabithorax, abdominal-A, Abdominal-B*) each of which specifies the characteristic development of a separate anatomical domain of the larva and the fly (Sanchez-Herrero *et al.*, 1985). Each of these three genes apparently encodes a protein containing a homeodomain (Regulski *et al.*, 1985).

Mutations of the *Ultrabithorax* (*Ubx*) function primarily affect the normal development of structures in posterior T2, in the whole of T3 and in anterior A1 (Morata and Kerridge, 1981; Struhl, 1984; Hayes *et al.*, 1984; Casanova *et al.*, 1985), i.e. in derivatives from parasegment 5 and 6 (Martinez-Arias and Lawrence, 1985). Some mutant effects can also be observed in abdominal segments (Lewis, 1978). In accordance with these genetic data is the early embryonic pattern of *Ubx* gene expression (Akam and Martinez-Arias, 1985). Accumulation of *Ubx* gene RNA is first observed at the blastoderm stage in parasegment 6. Soon thereafter, *Ubx* expression extends posteriorly through to parasegment 12 and appears at low levels in parasegment 5 and 13. The pattern of *Ubx* expression becomes increasingly complex in later development, although it remains most prominent in parasegment 6 (Akam, 1983; Akam and Martinez-Arias, 1985; White and Wilcox, 1984, 1985; Beachy *et al.*, 1985).

The *Ubx* genomic sequences have been cloned; the gene extends over more than one hundred kilobases (Bender *et al.*, 1983). The *Ubx* RNA products fall into three major size classes (Beachy *et al.*, 1985; Hogness *et al.*, 1985). The 4.3 kb and the 3.2 kb RNA contain two large protein coding exons as well as, in most cases, two miniexons; they probably differ from each other only with respect to their non-coding regions. The 4.7 kb RNA does not encode the large 3′ protein exon, is not found in the cytoplasm nor is it polyadenylated (Akam and Martinez-Arias, 1985), hence its function is unclear. The question arises whether these RNAs are initiated at the same promoter and thus subject to the same regulation modes.

The *Ubx* gene is activated in the early embryo either directly in response to positional information or indirectly, by the products of control genes activated prior to *Ubx* that establish early spatial organisation in the embryo. Characterisation of the regulatory elements in the *Ubx* promoter will be necessary for the identification of the proteins which regulate *Ubx* gene activation by binding to these elements. The regulatory proteins will provide a direct or indirect link to the molecular carrier of positional information. As a first step towards a *Ubx* promoter analysis we determined the sequence upstream of the N-terminal protein exon and mapped the initiation site of *Ubx* RNA within this region.

Results

The N-terminal *Ubx* protein exon (White and Wilcox, 1984) is encoded by sequences within an *Eco*RI/*Hin*dIII fragment of lambda phage L2229 (between map position −33/−31; Bender *et al.*, 1983). The protein initiator codon is located just upstream of the *Stu*I site within this fragment (Weinzierl *et al.*, 1987; Wilde and Akam, 1987). The *Eco*RI/*Stu*I fragment containing the non-coding sequences (611 bp) as well as the 5′ adjacent *Eco*RI subclone (3.5 kb insert) from phage L2229 (map position −31/−28) were used for sequence determination (Figure 1).

For S1 nuclease protection assays of RNA, we used various M13 subclones (see Figure 1). First, we used a single-stranded continuously labelled probe generated from mSE which includes sequences encoding the first eight amino acids of the *Ubx* protein. The mSE clone contains a third of the 'common probe' sequences used by Akam and Martinez-Arias (1985) to detect all three types of *Ubx* RNAs. We therefore assume that the mSE probe detects at the least all the *Ubx* protein-coding RNAs, but probably all *Ubx* RNA products. This probe was fully protected from S1 digestion by embryonic RNA, although not by RNA from Schneider cells (Figure 2a). Additional bands of lower molecular weight were observed. We believe that these do not represent sequence divergence points between genomic DNA and RNA, but that they indicate artifactual S1 cleavage at long runs of A's in the mRNA leader for the following reasons: their appearance is variable and dependent on the S1 digestion conditions (minimal at low temperature) and their size is consistent with S1 cleavage at position +466 and +752/+790 (Figure 1).

Second, a probe derived from the 5′ adjacent clone mEN gave rise to a protected fragment of 358–368 nt (Figure 2a). The signals obtained with both the mSE and the mEN probe were

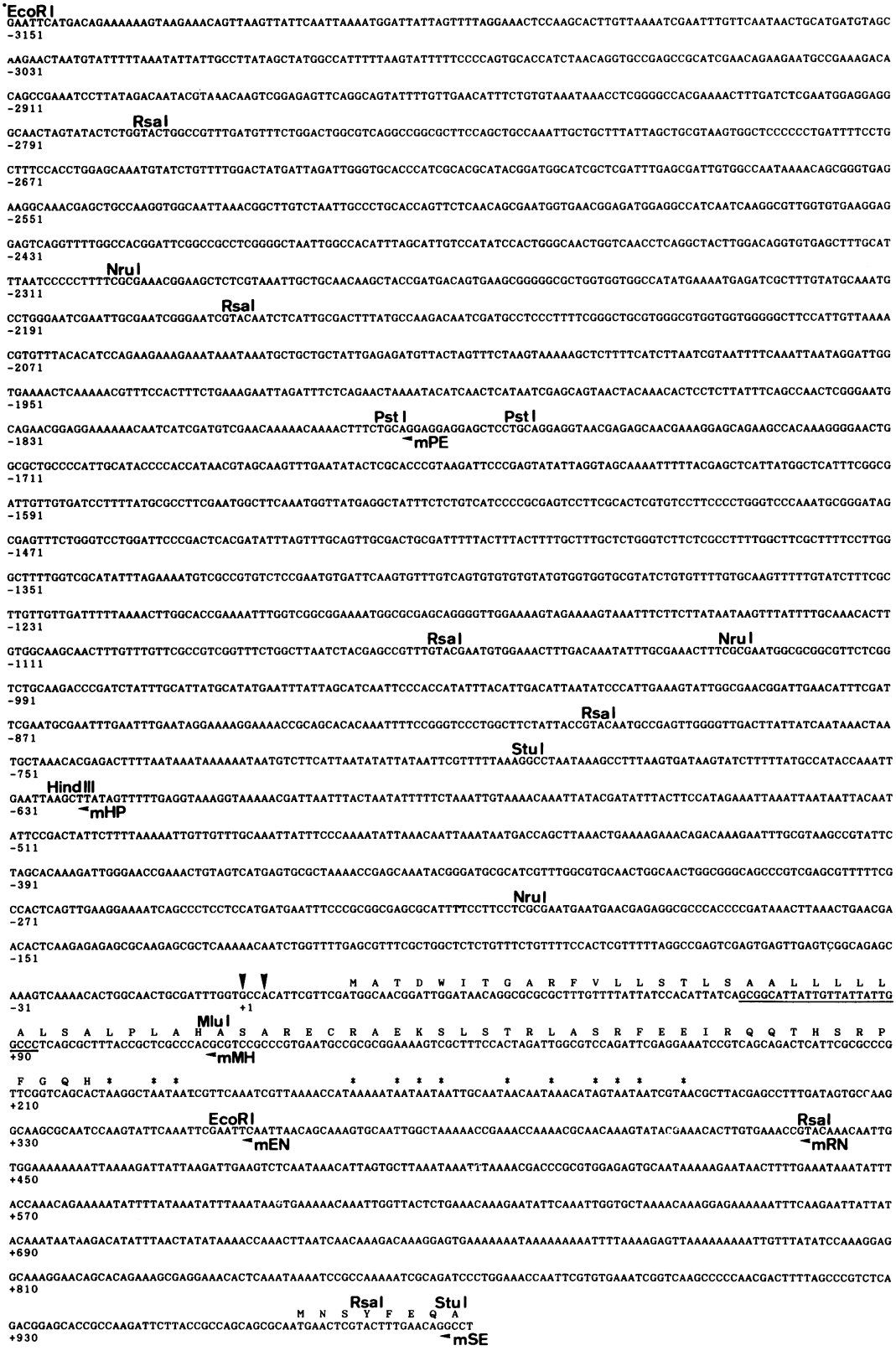

Fig. 1. Sequence of the *Ubx* promoter region. The sequence is divided into three parts: 3151 nt of 5′ flanking sequence, 967 nt of leader sequence and the first 25 nt of the *Ubx* protein coding sequence. Two adjacent transcriptional start sites are indicated by vertical arrows; the first of these residues (G) was termed +1. Various restriction enzyme cleavage sites are indicated (the most 5′ located *Eco*RI site is an artificial site corresponding to the end of the lambda clone L2229; Bender *et al.*, 1983). Arrows underneath some of the sites indicate the 3′ end of M13 subclones (mSE, mRN, mEN, mMH, mHP, mPE) that were used for S1 nuclease mapping of transcripts (see Materials and methods). Underlined is the sequence of the synthetic oligonucleotide which was used for primed extension. The long open reading frame in the leader (starting at +15) is translated into amino acids shown above the sequence. It terminates at a TAA stop codon at position +222 after a histidine codon; the following stop codons (11 TAA/1 TAG) are indicated by a star. The first eight amino acids of the *Ubx* N-terminal protein domain are shown at the bottom of the sequence.

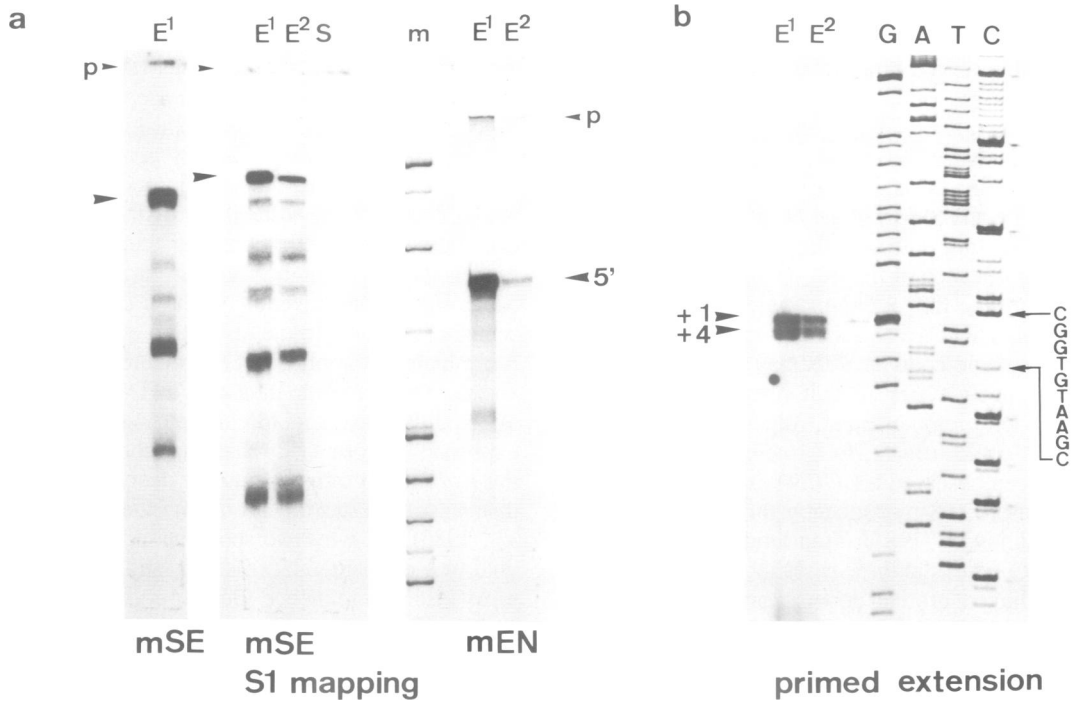

Fig. 2. 5′ end mapping of *Ubx* transcripts. (**a**) Total embryonic RNA (E^1: 2–9 h, E^2: 2–24 h after oviposition) or total RNA from *Drosophila* Schneider cells (S) was hybridised with a continuously labelled radioactive probe derived from mSE or mEN, and the fragments protected from S1 nuclease digestion were separated on a 7% sequencing gel. Bands corresponding to the sequence divergence point are marked by arrow; residual amounts of S1 resistant probe are marked with p (mSE ~940 nt, mEN 580 nt). The major band in the case of mSE reflects full length protection (size ~600 nt); two experiments are shown to indicate the variability of the lower bands (mainly 280 nt and 200 nt) which correspond to cleavage at long poly(A)/poly(T) stretches (see text). The major band in the case of mEN (358–368 nt; sized with respect to the bands in the marker lane, m) corresponds to the 5′ end of *Ubx* RNA (5′). *Ubx* RNA can clearly be detected in embryonic RNA preparations (the signal in the E^2 lane of this particular mEN mapping is unusually low), but not in *Drosophila* Schneider cells. (**b**) A primed extension experiment was done with embryonic RNA (see above) hybridised to a synthetic oligonucleotide derived from the *Ubx* leader (see Figure 1). A sequence reaction (same primer) is displayed next to it. The top band (+1) corresponds to a G residue (just above the double band in the adjacent C lane), the second band (+4) to an A residue. These two bands map to the same site as the protected band of the mEN probe (Figure 2a) and, therefore, this site represents the transcription initiation site for *Ubx* RNA.The first 10 residues of *Ubx* RNA are displayed on the right.

of similar intensity. We conclude that most or all *Ubx* RNA sequences diverge from the genomic sequences ~950 bp upstream of the methionine start codon. The sequence diversion point represents either the 5′ end or a splice acceptor site of *Ubx* RNA.

The position of the sequence divergence was confirmed by further mapping experiments using the mMH and mRN probes: these probes gave rise to protected fragments of ~120 nt and 440 nt, respectively (not shown). The latter probe served to ensure that we did not miss a small *Eco*RI fragment around map position −31 which could have been lost in the original subcloning from the lambda phage. The mHP and mPE probes containing sequences upstream of the putative *Ubx* 5′ end did not detect any RNA (not shown), indicating that there is no transcriptional activity on the same strand upstream of that putative 5′ end. Presumably, the closest transcription unit upstream of the *Ubx* transcription unit is located within the *bxd* region; *bxd* transcripts however terminate more than 7 kb upstream of the putative *Ubx* 5′ end (Hogness *et al.*, 1985).

An oligonucleotide derived from sequences located close to the putative *Ubx* 5′ end (see Figure 1) was used for a primed extension experiment (Figure 2b). The result clearly demonstrates that the sequence divergence point revealed by S1 mapping indeed represents an RNA initiation rather than a splice acceptor site: the main band obtained with primed extension corresponds to the G residue termed +1 (Figure 1). A second band just below the main band was observed (Figure 2b), indicating a second start site at the A residue at position +4. In summary, we conclude that probably all protein-coding RNAs from the *Ubx* gene originate at the same RNA initiation site. This implies a unique *Ubx* promoter that regulates expression of *Ubx* RNA.

Discussion

Our sequence and RNA mapping data predict (i) that *Ubx* mRNA has an unusually long mRNA leader and (ii) that, most likely, all *Ubx* mRNA initiate at the same site and are thus subject to control by a unique promoter. We would like to discuss the putative role of the *Ubx* RNA leader and 5′ flanking sequences in controlling *Ubx* expression.

The leader sequence contains only two methionine codons in addition to the *Ubx* protein initiator codon. The first of these methionine codons (at position +15) leads into an open reading frame of 204 nt; the second precedes a very short (different) open reading frame within this stretch of 204 nt. The first open reading frame is likely to be translated: the initiator codon context is favourable though not perfect for ribosome binding (Kozak, 1986) and the sequence within the open reading frame is highly non-random, a feature found almost exclusively in protein coding regions (positional base preferences in all three codon positions; Staden, 1984). The putative polypeptide contains a very hydrophobic N-terminal half (90% nonpolar amino acids, including serines) and a hydrophilic C-terminal half. The 23 contiguous nonpolar amino acids within the N-terminal part may indicate a trans-membrane domain. In a search through known protein

sequences, this N-terminal part was found to be homologous to N-terminal signal peptides of several mammalian transmembrane proteins as well as to mouse polyoma large T-antigen (43% identity to residues 417−439).

It is possible that this open reading frame in the *Ubx* mRNA leader serves some regulatory function, and the question arises whether it does so by acting in *cis* or in *trans*. The high sequence conservation observed in the 5′ part of the *D. funebris Ubx* leader (near-identity up to position +86; Wilde and Akam, 1987) indicates functional significance of this region. The conservation however includes only the first 23 amino acids of the putative leader peptide; termination of the *D. funebris* open reading frame occurs after 30 amino acid codons. This may suggest that it is not the *Ubx* leader peptide *per se* that provides a potential regulatory function. On the other hand, initiation at the leader peptide AUG codon could affect *Ubx* protein expression in a similar way as has been observed for *GCN4* expression in yeast (Müller and Hinnebusch, 1986) or recently for SV40 early protein expression (Khalili *et al.*, 1987). Translation initiation at the upstream leader AUG codons in these cases inhibits efficient expression of the *cis*-linked protein downstream. Open reading frames were found in mRNA leaders of other genes, notably in the *Drosophila fushi tarazu* gene (Laughon and Scott, 1984), and translational downregulation of *cis*-linked protein expression may be, at least in viruses, a common phenomenon (Khalili *et al.*, 1987). Finally, sequence conservation in the region of the open reading frame may reflect other regulatory functions: footprinting experiments with embryonic extracts reveal several protein binding sites (from +18 to +305) one of which strongly affects transcription *in vitro* if deleted (M.Biggin and R.Tjian, personal communication).

By analogy with other *Drosophila* and mammalian promoters it seems likely that the 3151 nt of 5′ flanking sequence described in this paper contain important transcriptional control elements: the majority of known enhancer and upstream elements are located within 3 kb of a gene (e.g. Serfling *et al.*, 1985; Garabedian *et al.*, 1986; Bienz and Pelham, 1987). However, in the case of *Ubx*, distant sequences as far as 30 kb upstream from the RNA initiation site apparently affect *Ubx* expression. Mutations which map within a region of 3−30 kb upstream of the *Ubx* transcription unit (*bxd, pbx*; Bender *et al.*, 1983) fail to complement *Ubx* mutations (Lewis, 1963). The main phenotypic defects observed in *bxd* and *pbx* mutations occur in derivatives of parasegment 6 (Lewis, 1963; Teugels and Ghysen, 1986). One of the most apparent effects of *bxd* mutations at the molecular level was found to be a reduced level of *Ubx* protein in parasegment 6 (Beachy *et al.*, 1985; Hogness *et al.*, 1985) where *Ubx* transcripts normally accumulate at high levels (Akam and Martinez-Arias, 1985). The overall spatial regulation of *Ubx* expression (parasegment 5−13) however remains unaffected in these *bxd* mutants. It was proposed that regulatory sequences within the *bxd* region act in *cis* on *Ubx* expression (Beachy *et al.*, 1985; Bender *et al.*, 1985; Hogness *et al.*, 1985).

We would like to suggest an alternative explanation: it is possible that the *novel* sequences brought into the vicinity of the *Ubx* promoter in the various *bxd* and *pbx* mutants (all of these are rearrangements or *gipsy* insertion mutants) negatively interfere with *Ubx* expression. This interference effect could qualitatively be the same in the various *bxd* mutants, although the extent of interference might vary among them. The same interference effect cannot completely account for the *pbx* mutants which show a phenotype distinct from the *bxd* phenotype; however, *pbx* mutations are generated by rare rearrangement events and may represent special, selected cases. Transcriptional interference as a mechanism for gene regulation has recently been demonstrated experimentally with globin genes (Proudfoot, 1986) and may naturally occur in one of the thalassaemias (Kioussis *et al.*, 1983). Moreover, it may be the mechanism by which *gipsy* elements reduce transcript levels of nearby genes (Parkhurst and Corces, 1985). It implies that transcription complexes formed at a particular promoter are destabilised by incoming RNA polymerases originating from (novel) upstream sequences. In contrast to the models mentioned in the previous paragraph, our explanation attributes the *cis*-inactivation effect on *Ubx* expression to the *novel* sequences and not to the disruption of *resident* sequences in the *bxd* mutants. It implies that *bxd* mutations should affect the level and not the spatial domain of *Ubx* expression and that the area affected the most should correlate wih the area of highest *Ubx* expression. More easily than the other models, it explains why the *bxd* mutant effect gradually decreases with increasing map distance of the *bxd* mutation from the *Ubx* promoter (Bender *et al.*, 1985) and why two spontaneous revertants of bxd^1 can be wild-type despite the fact that they still contain part of the original *gipsy* insertion sequence (Bender *et al.*, 1983). In summary, the resident *bxd* sequences may not be part of the *Ubx* promoter.

We have sequenced a region of the *Ubx* promoter which presumably contains target sites for regulatory proteins 'reading' positional information in the early embryo. The question arises whether these sequences confer transcriptional activity in an expression assay. Indeed, strong promoter activity is observed in an embryonic *in vitro* transcription system (M.Biggin and R.Tjian, personal communication). Preliminary results obtained with constructs containing all the above *Ubx* sequence fused to a beta-galactosidase protein indicate their promoter activity in stably transformed embryos; the expression pattern in these embryos shows some of the regulatory features of the resident *Ubx* gene (Bienz *et al.*, in preparation). However, the same constructs were not expressed in transiently transfected cells of three different *Drosophila* cell lines (G.Saari and M.Bienz, unpublished), as may be expected from the fact that the endogenous *Ubx* gene is not expressed in these cells (see Figure 2a). The transcriptional activity of the *Ubx* promoter sequences *in vitro* and in transformed embryos will allow functional analysis and dissection of the regulatory elements.

Materials and methods

Subcloning and sequence determination

Two *Eco*RI subclones from lambda phage L2229 (Bender *et al.*, 1983), pϕDm3102 and pϕDm3108 (Akam, 1983), were kindly provided by M.Akam. The following fragments were subcloned into M13 to generate clones suitable for sequencing and nuclease S1 mapping: a *StuI/Eco*RI 630 bp fragment from pϕDm3108 into *SmaI/Eco*RI cut mp8 (mSE): an *Eco*RI/*NruI* 560 bp fragment from pϕDm3102 into *Eco*RI/*Sma*I cut mp9 (mEN); an *MluI/Hin*dIII 740 bp fragment from pϕDm3102 into *SmaI/Hin*dIII cut mp9 (mMH); a *Hin*dIII/*PstI* 1140 bp fragment from pϕDm3102 into mp8 (mHP); a *PstI/Eco*RI 1380 bp fragment from pϕDm3102 into mp8 (mPE); an *RsaI/NruI* 640 bp fragment from a subclone containing pϕDm3102 joined to pϕDm3108 into *Sma*I cut mp8 (mRN). Additional M13 subclones were generated for determination of the complete sequence (see Figure 1) on both strands (Bankier and Barrell, 1983).

Transcript mapping

Total RNA was extracted from *Drosophila* embryos 2−9 h (E^1) or 2−24 h (E^2) after oviposition or from *Drosophila* Schneider cell lines (S) as previously described (Riddihough and Pelham, 1986). For nuclease S1 mapping, the procedure of Pelham (1982) was followed: 30 μg of total RNA after LiCl-precipitation was hybridised overnight in 80% formamide/400 mM NaCl/10 mM Pipes pH 6.5/1 mM EDTA at 50°C. When a probe derived from the clone mSE was used, S1 nuclease digestion was done for 30 min at 18°C (instead of 10 min at 37°C); full protection of this probe by RNA was only obtained at low temperature where artifactual cleavage of the RNA−DNA hybrids at long poly(A)/poly(T)

stretches was minimal. For primed extension experiments, the protocol of Jones *et al.* (1985) was used. The primer (25 mer, residues +69/+93 in Figure 1) was a gift from M.Biggin.

Acknowledgements

We thank Mark Biggin, Gines Morata and Hugh Pelham for discussion and helpful comments on the manuscript, Michael Akam for providing plasmids, Mark Biggin for providing a synthetic oligomer and the Swiss National Science Foundation (grant nr. 3.313-0.86) for financial support.

References

Akam,M.E. (1983) *EMBO J.*, *32*, 2075–2084.
Akam,M.E. and Martinez-Arias,A. (1985) *EMBO J.*, **4**, 1689–1700.
Bankier,A.T. and Barrell,B.G. (1983) In Flavell,R.A. (ed.), *Techniques in the Life Sciences*. Elsevier Scientific Publishers Ireland Ltd, Vol. B5, pp. 1–34.
Beachy,P.A., Helfand,S.L. and Hogness,D.S. (1985) *Nature*, **313**, 545–551.
Bender,W., Akam,M., Karch,F., Beachy,P.A., Pfeifer,M., Spierer,P., Lewis,E.B. and Hogness,D.S. (1983) *Science*, **221**, 23–29.
Bender,W., Weiffenbach,B., Karch,F. and Pfeifer,M. (1985) *Cold Spring Harbor Symp. Quant. Biol.*, **L**, 173–180.
Bienz,M. and Pelham,H.R.B. (1987) *Adv. Genet.*, in press.
Casanova,J., Sanchez-Herrero,E. and Morata,G. (1985) *Cell*, **42**, 663–669.
Garabedian,M.J., Shepherd,B.M. and Wensink,P.C. (1986) *Cell*, **45**, 859–867.
Hayes,P.H., Sato,T. and Denell,R.E. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 545–549.
Hogness,D.S., Lipshitz,H.D., Beachy,P.A., Peattie,D.A., Saint,R.B., Goldschmidt-Clermont,M., Harte,P.J., Gavis,E.R. and Helfand,S.L. (1985) *Cold Spring Harbor Symp. Quant. Biol.*, **L**, 181–194.
Jones,K.A., Yamamoto,K.R. and Tjian,R. (1985) *Cell*, **42**, 559–572.
Khalili,K., Brady,J. and Khoury,G. (1987) *Cell*, **48**, 639–645.
Kioussis,D., Vanin,E., deLange,T., Flavell,R.A. and Grosveld,F.G. (1983) *Nature*, **306**, 662–666.
Kozak,M. (1986) *Cell*, **44**, 283–292.
Laughon,A. and Scott,M.P. (1984) *Nature*, **310**, 25–31.
Lawrence,P.A. and Morata,G. (1983) *Cell*, **35**, 595–601.
Lewis,E.B. (1963) *Am. Zool.*, **3**, 33–56.
Lewis,E.B. (1978) *Nature*, **276**, 565–570.
Martinez-Arias,M. and Lawrence,P.A. (1985) *Nature*, **313**, 639–642.
Morata,G. and Kerridge,S. (1981) *Nature*, **290**, 778–781.
Müller,P.P. and Hinnebusch,A.G. (1986) *Cell*, **45**, 201–207.
Parkhurst,S.M. and Corces,V.G. (1985) *Cell*, **41**, 429–437.
Pelham,H.R.B. (1982) *Cell*, **30**, 517–528.
Proudfoot,N.J. (1986) *Nature*, **322**, 562–565.
Regulski,M., Harding,K., Kostriken,R., Karch,F., Levine,M. and McGinnis,W. (1985) *Cell*, **43**, 71–80.
Riddihough,H. and Pelham,H.R.B. (1986) *EMBO J.*, **51**, 1653–1658.
Sanchez-Herrero,E., Vernos,I., Marco,R. and Morata,G. (1985) *Nature*, **313**, 108–113.
Serfling,E., Jasin,M. and Schaffner,W. (1985) *Trends Genet.*, **1**, 224–230.
Staden,R. (1984) *Nucleic Acids Res.*, **12**, 551–567.
Struhl,G. (1984) *Nature*, **308**, 454–457.
Teugels,E. and Ghysen,A. (1986) *Nature*, **314**, 558–561.
Weinzierl,R.O.J., Axton,M., Ghysen,A. and Akam,M. (1987) *Genes and Dev.*, in press.
White,R.A.H. and Wilcox,M. (1984) *Cell*, **39**, 163–171.
White,R.A.H. and Wilcox,M. (1985) *EMBO J.*, **4**, 2035–2043.
Wilde,C.D. and Akam,M. (1987) *EMBO J.*, 1392–1401.

Received on February 25, 1987; revised on April 1, 1987