

# Properties of promoter regions of *mdg1* *Drosophila* retrotransposon indicate that it belongs to a specific class of promoters

Irina R. Arkhipova<sup>1,2,3</sup> and Yurii V. Ilyin<sup>1</sup>

<sup>1</sup>V.A. Engelhardt Institute of Molecular Biology, Academy of Sciences of the USSR, Vavilov str. 32, 117984 Moscow, USSR and <sup>2</sup>Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh EH9 3JR, UK.

<sup>3</sup>Reprint requests to: Department of Biochemistry and Molecular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138, USA.

Communicated by G.P. Georgiev

**A sequence 30 bp downstream from the start site of the *Drosophila melanogaster* retrotransposon *mdg1* is shown to be responsible for correct and precise initiation of *mdg1* RNA synthesis in combination with the RNA start-site sequence TCAGTT. A sequence-specific DNA binding protein is demonstrated to interact with the +30 sequence, and the efficient binding of this factor is necessary for *in vivo* transcriptional activity of the plasmid constructs containing *mdg1* promoter fragments. The nucleotides –8/+34 of *mdg1* represent a minimal promoter which is able to provide correct initiation of transcription by RNA polymerase II at basal levels. A comparison with properties of some other retrotransposable elements and several developmentally regulated cellular genes allows us to conclude that together they form a specific class of RNA polymerase II promoter. This promoter class characteristically lacks upstream sequences necessary for transcription initiation, such as TATA boxes, but requires a specific downstream promoter element within 40 bp downstream of the RNA start site. The level of transcription can, however, be modulated by upstream regulatory elements. The identified sequence-specific downstream initiation factor may be responsible for transcription initiation on promoters of some genes which belong to this class.**

**Key words:** *Drosophila* retrotransposons/downstream initiation factor/RNA polymerase II promoter/transcription initiation

## Introduction

*Drosophila* retrotransposons are representatives of a wide class of eukaryotic mobile genetic elements similar to vertebrate retroviruses in their structure and replication cycle (Georgiev, 1984; Finnegan and Fawcett, 1986; Boeke, 1988, 1989; Bingham and Zachar, 1989; Boeke and Corces, 1989). They are transposed via the reverse transcription of their corresponding RNAs (Boeke *et al.*, 1985; Arkhipova *et al.*, 1984, 1986) and active copies are transcribed at different developmental stages as well as in cultured cells. In general, these elements do not encode their own transcription factors, but use the cellular RNA polymerase II transcription system. For several *Drosophila* and yeast retrotransposons, a set of cellular regulatory genes involved in their transcription has

been identified and characterized (Boeke, 1989; Mazo *et al.*, 1989; Spana *et al.*, 1988). These genes encode transcriptional activators or repressors able to bind to certain *cis*-acting regulatory regions within the retrotransposon body. The transcriptional control sequences within the long terminal repeats (LTRs), where transcription is initiated and terminated, are not well defined.

The general design of typical RNA polymerase II (polII) promoters is rather complex (Shenk, 1981; Struhl, 1987; Wasylyk, 1988; Dynan, 1989). There are a number of discrete sequence elements which may be dispersed at various distances upstream or downstream from the RNA start sites and which bind numerous protein factors. *Cis*-acting sequences necessary for transcription initiation and modulation are often considered as promoters themselves and consist of the start-site selection sequences and upstream promoter elements; more distal, orientation- and often position-independent regulatory elements are usually called enhancers. The most severe positional constraints are placed upon the start-site selection sequences including a highly conserved TATA-box sequence located ~25–30 bp upstream from the RNA start site. This element appears to be responsible for correct and precise initiation of mRNA synthesis, since mutations in it lead to heterogeneity of RNA start sites (e.g. Mathis and Chambon, 1981). Usually the –40 to +30 bp sequence is considered to be the interaction site for general transcription factors able to interact with the majority of polII promoters. Some promoters, such as those for many housekeeping genes, lack discernible TATA motifs, but they contain very GC-rich regions in this area (Dynan, 1986).

In an earlier study, while determining the position of RNA start sites within the LTRs of three *D. melanogaster* retrotransposons (*mdg1*, *mdg3* and *gypsy*, or *mdg4*), homogeneous RNA start sites were observed for all three elements, but none of them contained any sequences resembling TATA boxes at the appropriate position in the 5'-nontranscribed regions (Arkhipova *et al.*, 1986). Nor did they possess GC-rich regions typical of the non-TATA promoters of housekeeping genes. Moreover, all of the three different elements had a common sequence, TCAGTPy, immediately adjacent to the RNA start site (Figure 1). Such an unusual organization of the *mdg* promoter region stimulated us to determine the LTR sequences necessary for efficient and precise initiation of RNA synthesis.

Traditionally, a minimal promoter is considered to be represented by the TATA element in conjunction with the cap site (for review, see Lewin, 1990). The only component of the general polII transcription machinery known to possess a sequence-specific DNA binding activity is the factor TFIID which binds specifically to the TATA motif. However, the basis for transcription from a number of TATA-less promoters (for references, see Dynan, 1989; Smale and Baltimore, 1989) has not yet been identified. This study

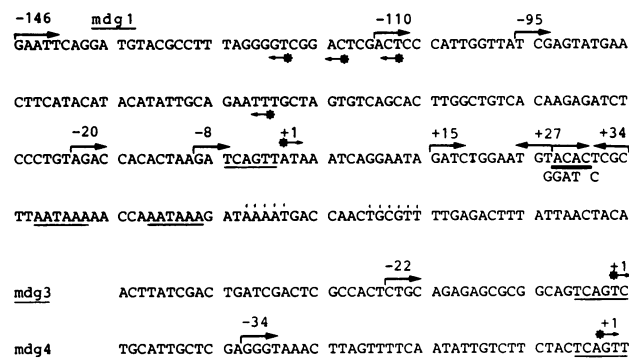
represents the first evidence of a downstream sequence-specific DNA binding protein which is necessary for transcription initiation on a minimal promoter consisting of the RNA start site and the downstream +30 element.

## Results

### *mdg1-CAT constructs and their expression*

To identify the minimal promoter sequences, we created a series of plasmid constructs which contained a standard reporter bacterial chloramphenicol acetyltransferase (CAT) gene followed by SV40 intron/poly(A) site (Gorman *et al.*, 1982) under the control of the LTR sequences deleted to various extents (see Materials and methods). The most detailed study was performed for *mdg1*. Figure 2 is a schematic representation of *mdg1*-CAT constructs containing 5'- and 3'-deletions of the LTR (see also Figure 1), as well as insertions and inversions. These constructs were tested in transient expression assays in the *D.melanogaster* Schneider 2 cell line.

Transfections and CAT assays performed using this series of constructs revealed the following picture of CAT



**Fig. 1.** Nucleotide sequences of the LTR regions adjacent to the sites of initiation of RNA synthesis for retrotransposons *mdg1*, *mdg3* and *gypsy* (*mdg4*). The transcriptional start sites and polyadenylation sites determined by Arkipova *et al.* (1986) are indicated by arrows with asterisks and by dots, respectively. The sequence TCAGT<sub>C</sub> and the polyadenylation signals are underlined. 5'- and 3'-deletion endpoints are shown by horizontal arrows. Arrowed asterisks in an opposite orientation denote RNA start sites from the outward *mdg1* promoter determined in the present work. The letters below the main sequence indicate base changes in the +30 core (underlined by a thick line) mutant clones.

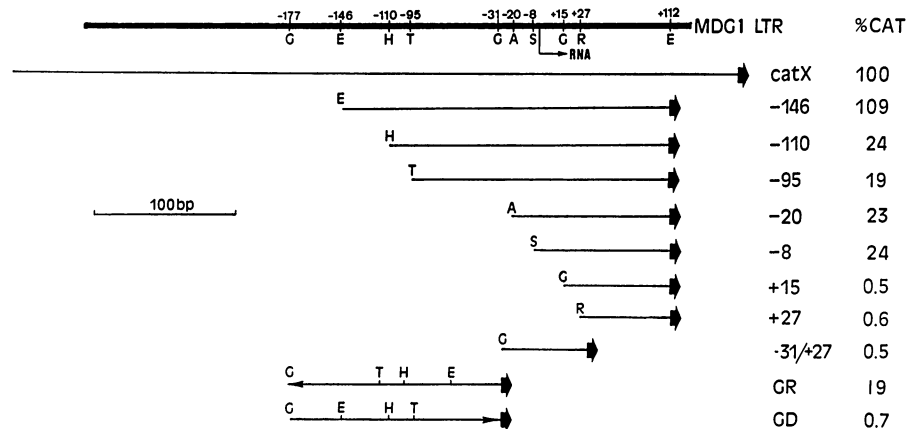
expression (indicated in Figure 2). The entire LTR (*catX*), as well as the -146 bp deletion mutant, is able to direct high levels of CAT enzyme activity. Further 5'-deletion constructs exhibit a several-fold decrease in CAT activity, indicating the presence of an essential upstream distal promoter element located between the *EcoRI* and *HinI* sites (-146 to -110 bp). The sequence of this region does not reveal any homologies to known transcription factor binding sites, although there is a weak homology with the upstream regulatory element of the *copia* retrotransposon activating its transcription [-133(ACGCCTTTAGG)-123 for *mdg1* and -60(ACACCTTTTATC)-50 for *copia*] (Sneddon and Flavell, 1989). This upstream promoter element probably cannot act as an enhancer, since the fragment containing this element does not produce any effect on transcription when inserted upstream from the *mdg4* promoter at the same distance from the RNA start site (data not shown).

The *mdg1* sequence (-179/-31) upstream from the RNA start site does not possess promoter activity when inserted into the vector pUSVL-*cat* in direct orientation (construct GD). However, the same sequence inserted in the opposite orientation (construct GR) can produce measurable levels of CAT activity, thus explaining the phenomenon of bidirectional transcription of some *mdg* elements observed in earlier studies (Ilyin *et al.*, 1980). In mobile elements, a bidirectional promoter has also been found in the *Lepidoptera* retrotransposon TED (Friesen *et al.*, 1986).

The most surprising result was that all the further 5'-deletions up to the position -8 did not cause a significant decrease in CAT activity. Only the deletions extending into the transcribed region (positions +15 and +27) reduced activity to the background level of the promoterless vector pUSVL-*cat*. Deletion of the sequence from the other direction (3' end of the LTR) also caused a drop in activity of the promoter thus indicating the presence of important downstream elements in the transcribed region.

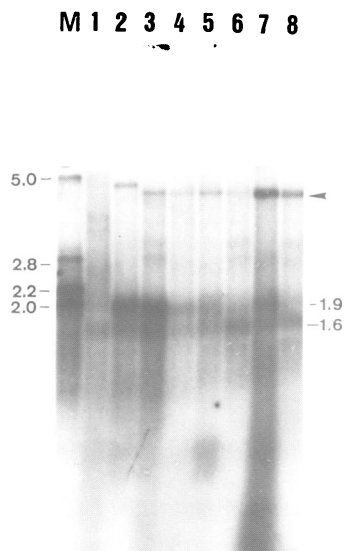
### *Northern blot analysis*

To exclude the possibility of changes at the level of translation, we performed Northern blotting experiments with poly(A)<sup>+</sup> RNA extracted from culture cells transfected with some of the constructs. The results of these experiments are presented in Figure 3, which demonstrates that the data on the CAT enzyme expression are in good agreement with



**Fig. 2.** Schematic representation of *mdg1*-CAT fusions. The fused CAT sequences are shown by thick arrows. The names of the clones are given in the left column; the figures in the right column denote the mean level of their CAT activity with respect to the longest construct. Restriction endonucleases are abbreviated as follows: A, *AccI*; G, *BglII*; E, *EcoRI*; H, *HinI*; R, *RsaI*; S, *Sau3A*; T, *TaqI*.

the Northern analysis. mdg1-CAT transcripts of the correct length correspond to the upper of RNA bands (1.9 kb), while the lower 1.6 kb band is most probably initiated within the CAT gene. The highest RNA levels for the 1.9 kb band are observed for the constructs catX and -146 (lanes 2 and 3). The intensity of this band is reduced significantly for the constructs -110, -20 and -8 (lanes 4, 5 and 7), and it



**Fig. 3.** Northern blot analysis of RNAs expressed in transient transfection experiments. The  $^{32}\text{P}$ -labelled *Hin*I fragment of the vector pUSVL-cat was used as a probe. Transcription of the following constructs is shown: GR (1), catX (2), -146 (3), -110 (4), -95 (5), +15 (6), -8 (7), +27 (8). M denotes the lane with restriction fragments of the DNA marker plasmid hybridizable to the CAT probe; the size of the marker fragments is indicated on the left. The arrow at the slowly migrating bands indicates contaminating plasmid DNA used for transfection in large quantities (25  $\mu\text{g}/10$  ml plate).

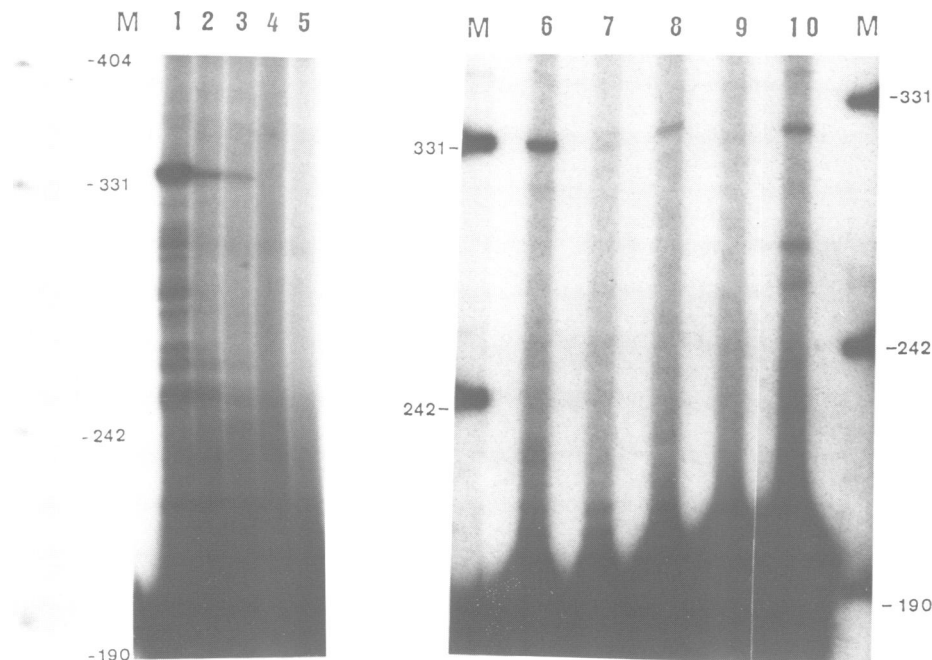
is not visible for the constructs +15 and +27 (lanes 6 and 8) which produce only the shorter non-mdg1-specific 1.6 kb transcript.

Two other retrotransposons, mdg3 and mdg4 (gypsy), were examined in less detail. mdg3-CAT fusions displayed no difference in expression from the constructs -109/+47 and -22/+47, the same being true at the RNA level, and gypsy was equally well expressed in the -35/+192 construct when compared with the entire LTR (data not shown). However, from these constructs it cannot be excluded that regulatory sequences are situated nearer than -22 bp for mdg3 and -35 bp for mdg4.

#### Primer extension experiments

To confirm that the RNA synthesized from construct -8 was not initiated at cryptic downstream or upstream sites and that its 5' end remained homogeneous, we performed primer extension experiments with the transiently expressed RNAs. The results are presented in Figure 4. It may be seen that the RNA synthesized from construct -8 (lane 2), as well as from -20 (lane 8), preserves the correct and highly homogenous start site when compared with the construct -146 (lanes 1 and 6). The length of the extended primer corresponds to the previously defined RNA start site for endogenous genomic mdg1 copies (Arkhipova *et al.*, 1986). As in the Northern blotting experiments, RNA synthesis is not observed for the constructs +15 and +27 or for the vector plasmid (lanes 4, 7 and 5); nor is it observed with non-promoter-containing inserts (lane 9).

The same RNA start site is also observed for the -8 construct in transfected culture cells of *Drosophila hydei* which do not carry endogenous mdg1 copies (lane 3). This indicates that the general transcription factors responsible for its transcription are well conserved between *D. melanogaster* and *D. hydei*. It is interesting that there is some block of mdg1-CAT translation in *D. hydei* cells, since the

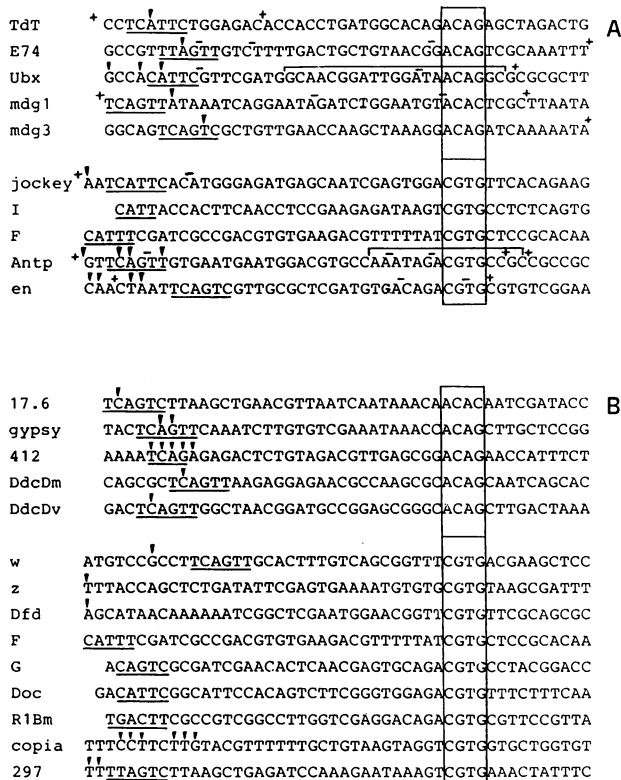


**Fig. 4.** Primer extension analysis of transiently expressed mdg1-CAT constructs using the  $^{32}\text{P}$ -labelled primer synthesized on M13cat single-stranded template. The following plasmids were tested: 1, 6, -146; 2, -8; 3, -8 in *D. hydei*; 4, +15; 5, vector pUSVL-cat; 7, +27; 8, -20; 9, non-promoter-containing fragment of *jockey* (Priimagi *et al.*, 1988); 10, GR. The marker used was the pUC19 plasmid cleaved by *Msp*I.

level of CAT expression for several constructs which produce enough RNA (judged by Northern blotting, data not shown) is almost undetectable.

The LTR fragment inserted in the opposite orientation, which exhibited an outward promoter activity (construct GR), was shown in these experiments to direct several RNA start sites (Figure 4, lane 10; also shown in Figure 1).

The construct  $-8$  retains upstream of the transcriptional start site only the sequence TCAGTT which is thought to be important as a strong RNA start site (see Discussion). Therefore, the sequences responsible for *mdg1* LTR promoter activity must lie in the region downstream from the site of transcription initiation. The transcribed region, after the first 35 bp, contains a long 30 bp very AT-rich sequence which includes two successive polyadenylation

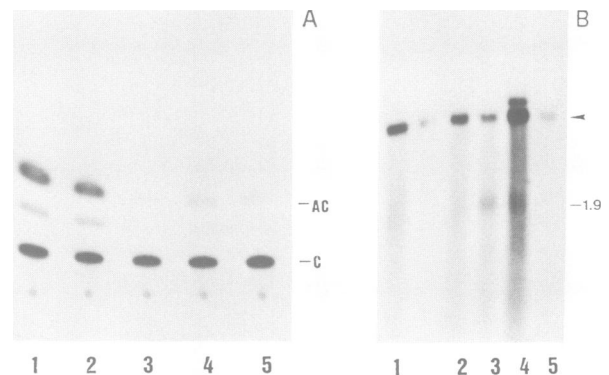


**Fig. 5.** Nucleotide sequence of the transcribed downstream region of *D.melanogaster* retrotransposons *mdg1*, *mdg3*, *gypsy* (*mdg4*) (Arkipova *et al.*, 1986), 412 (Yuki *et al.*, 1986), 17.6 (Inouye *et al.*, 1986), 297 (Matsuo *et al.*, 1986), *copia* (Emori *et al.*, 1985); LINE elements of *D.melanogaster* *jockey* (J1, Mizrokhi *et al.*, 1988), I factor (IR1, Fawcett *et al.*, 1986), F element (101F, DiNocera *et al.*, 1983), G element (DiNocera, 1988), Doc (Driver *et al.*, 1989); of *Bombyx mori* R1Bm (Xiong and Eickbush, 1988), *D.melanogaster* developmentally regulated genes *Ubx* (Biggin and Tjian, 1988), *en* (Soeller *et al.*, 1988), *Antp* (P2, Perkins *et al.*, 1988), *Dfd* (Regulski *et al.*, 1987), *z* (Pirrota *et al.*, 1987), *w* (O'Hare *et al.*, 1984), E74 (Thummel, 1989), *D.melanogaster* and *D.virilis* *Ddc* (Bray and Hirsh, 1986; Hirsh *et al.*, 1986); and mammalian lymphocyte-specific TdT (Smale and Baltimore, 1989). The sites of initiation of RNA synthesis (where known) are indicated by vertical arrows, the sequences TCAGTY or similar are underlined. Core regions of homology for two groups are boxed. (In principle, the *en* sequence can also be ascribed to the first group since both types of cores are adjacent and conserved between *D.melanogaster* and *D.virilis* and there are two neighbouring clusters of RNA start sites.) Square brackets denote the footprinted regions of *Ubx* and *Antp* (see references above). Pluses and minuses in the 3' part of the sequences designate 3'-deletion mutants which are either transcribed or not; those in the 5' part designate respective 5'-deletion mutants.

signals and polyadenylation sites (Figure 1). The first 35 bp, therefore, are the most likely candidates to be responsible for the promoter activity. Then we used the *RsaI* site at position +27 to create the 3'-deletion, and removal of the sequences downstream from +27 resulted in loss of CAT expression and of correctly initiated RNA synthesis (Figure 2). Therefore, an important promoter element must be removed by this deletion.

#### Nucleotide sequence comparisons

It was of interest to compare the first 50 bp after the transcriptional start between different genes proven and thought to have downstream promoter sequences. We have analysed retrotransposons for which the positions of RNA start sites have been determined (*mdg1*, *mdg3*, *gypsy*, 412, *copia*, 17.6, 297), several LINE elements from *Drosophila* and other species with confirmed (*jockey*) or possible (I, F, G, Doc, R1Bm) RNA polIII internal promoters, and several *Drosophila* and mammalian non-TATA genes which are expressed in a regulated fashion and known to have transcriptionally important downstream elements [*Ultrabithorax* (*Ubx*), *engrailed* (*en*), *Antennapedia* (*Antp*), ecdysone-responsive E74 gene, lymphocyte-specific terminal deoxynucleotidyltransferase (TdT)] and some other developmentally regulated non-TATA *Drosophila* genes [*white* (*w*), *zeste* (*z*), *Deformed* (*Dfd*), and *D.melanogaster* and *D.virilis* dopa decarboxylase (*Ddc*) genes which have a TATA sequence which is not important for correct initiation of transcription] (for references, see the legend to Figure 6). The sequence similarities between LINEs, *Antp* and *en* have already been noticed (C.McLean and D.Finnegan, personal communication). A small conserved core sequence can easily be distinguished in all these genes in the region 30–40 bp downstream from the RNA start site (Figure 5). Panel A represents the promoter sequences for which functional analysis has been carried out. Panel B



**Fig. 6.** (A) CAT activity of the transfected plasmids  $-8/+34$  (1),  $-8/+112$  (2),  $-8/+34M$  (3),  $-8/+112M$  (4) and pUSVL-cat (5). The volume of the extracts used for CAT assays was normalized according to  $\beta$ -galactosidase activity of the cotransfected D88 plasmid containing the *lacZ* gene (see Materials and methods). (B) Northern blot analysis of poly(A)<sup>+</sup> RNA extracted from Schneider 2 tissue culture cells transfected with pUSVL-cat (1),  $-8/+34M$  (2),  $-8/+34$  (3),  $-8/+112$  (4) and  $-8/+112M$  (5). The hybridization probe was the same as in Figure 3. The arrow again indicates contaminating transfected plasmid DNA; the additional 1.6 kb transcript coming from the vector in Figure 3 is not observed in this particular Schneider 2 cell line.

represents alignment of corresponding sequences of several *Drosophila* genes which may be expected to have downstream promoter elements, being LINE-like elements, or non-TATA retrotransposons, or developmentally regulated non-TATA genes. Interestingly, two types of the core consensus can be distinguished, either ACAG/C or CGTG. It is of particular interest that retrotransposons 17.6 and 297, which are highly homologous and considered to have a common origin, fall into different groups. It is also worth mentioning that comparison of corresponding regions in different *Drosophila* species [*Ddc* (Bray and Hirsh, 1986) and *en* (Kassis *et al.*, 1989) from *D.melanogaster* and *D.virilis*, as well as *Ubx* from *D.melanogaster* and *D.funnebris* (Wilde and Akam, 1987)], an approach which usually reveals islands of homologous sequences having functional significance in the non-coding region of the genes, shows that the sequences in the region in question represent blocks of conservation.

In all the cases where functional studies involving 5'- and 3'-deletions were performed (*mdg1*, *Ubx*, *en*, *Antp*, *E74*; present work and the references in the legend to Figure 6), precise removal of the core sequence by 3'-deletion (deletion endpoints are indicated in Figure 6) abolished correct transcription initiation, and the same was true for 5'-deletions removing the RNA start-site sequence. Note that for TdT gene the correct RNA start at low levels was preserved in the +11 3'-deletion; note also that the sequence upstream of +11 has the element RRAGACA which is present in the downstream core; point mutation in the first T of the start-site sequence TCATT drastically reduced transcription (Smale and Baltimore, 1989). In two cases (*Ubx*, *Antp*) protein-binding sites have been observed as DNase I footprints of downstream regions. Thus, the identified core sequences are very likely to be functionally significant, and at least two important elements (RNA start-site sequence and the downstream element) appear to constitute a minimal promoter.

#### ***mdg1* transcription is dependent on the +30 core**

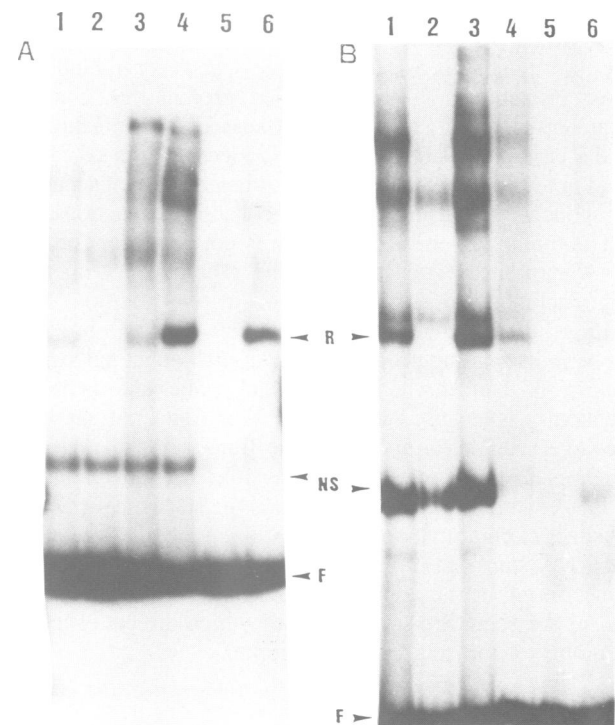
To investigate further the functional significance of the identified +30 core, several other constructs were created by cloning of the synthetic oligonucleotides containing wild-type and mutant *mdg1* promoter sequences into the same vector. The first construct (-8/+34) represented the wild-type *mdg1* promoter, the second (-8/+34M) carried a 5 bp mutation in the +30 core (TACACT → GGATCC), and the third (-8/+112M) was derived from -8/+112 but had the 21mer carrying the mutant sequence inserted between the RNA start site and the wild-type +30 core at position +15 in the same orientation (for details see Materials and methods and Figure 1). Transcriptional activity of these promoter fragments is represented in Figure 6. CAT assays (Figure 6A) demonstrate that the activity of the construct -8/+34 is approximately equal to that of the previously investigated -8/+112 construct (lanes 1 and 2), while the activity of the mutant clone -8/+34M and of the clone -8/+112M (in which the wild-type +30 core is removed further 21 bp downstream from the RNA start site upon insertion of the mutant oligonucleotide) is comparable to that of the vector plasmid (lanes 3-5). Thus, all the sequences necessary for basal level of *mdg1* transcription are located in the -8/+34 region.

Analysis of RNA extracted from transfected Schneider 2

tissue culture cells (Figure 6B) shows that, in agreement with the data on CAT activity, detectable transcripts of appropriate length can be seen only for the wild-type (lanes 3 and 4), but not the mutant, -8/+34 and -8/+112 constructs. One can conclude from these experiments that the +30 core ACAC does play an important role in *mdg1* expression at the transcriptional level and that the change in spacing (insertion of 21 bp) between the RNA start site and the +30 core also produces dramatic effects on the transcriptional activity in the *in vivo* assays.

#### **A protein factor binds sequence-specifically to the +30 core**

The above studies suggested the existence of a protein factor, binding of which to the downstream +30 core is required for transcription initiation by polIII. To study DNA-protein interactions in the promoter region, nuclear extracts from Schneider 2 cells were tested for the presence of a sequence-specific DNA binding activity dependent on the +30 core. As probes for gel mobility shift experiments, <sup>32</sup>P-end-labelled promoter-containing *mdg1* fragments of convenient length (~200 bp) were used. Figure 7A shows that a sequence-specific retarded band is observed with the probe derived from the minimal promoter (-8/+34) construct (lane 4): this band can be competed by the wild-type (lane 5), but not the mutant (lane 6) oligonucleotide. The fastest migrating complex (NS) in Figure 7 never exhibited a clear-cut sequence specificity in DNA binding and was further



**Fig. 7.** Gel mobility shift assays of the nuclear extracts isolated from Schneider 2 cultured cells. As probes, the following <sup>32</sup>P-end-labelled *mdg1* promoter fragments were used: (A) 200 bp *HindIII*-*PvuII* fragment of -8/+34M (lanes 1-3) and -8/+34 (lanes 4-6); (B) 150 bp *HindIII*-*EcoRI* fragments of -8/+112 (lanes 1-3) and -8/+112M (lanes 4-6). The wild-type and mutant *mdg1* oligonucleotides (see Materials and methods) were used as competitors in lanes 2, 5 (wild-type) and 3, 6 (mutant). R denotes the sequence-specific retarded band; NS, nonspecific complex; F, free probe.

regarded as a non-specific DNA–protein complex. The probe derived from the mutant  $-8/+34M$  construct (lanes 1–3) exhibits essentially the same gel retardation pattern and analogous competition properties (i.e. the retarded band can be competed by the wild-type but not by the mutant oligonucleotide), but its intensity is reduced dramatically ( $\sim 100$ -fold). These data can be interpreted in terms of protein binding with little or no sequence specificity to the mutant promoter probe, which can be greatly enhanced by the presence of the wild-type  $+30$  core sequence representing an efficient binding site. The efficient binding of this factor correlates with the transcriptional activity of the construct (see previous section).

The probe representing *mdg1* promoter sequence including more downstream regions  $-8/+112$  (Figure 7B) reveals a more complex picture of retarded bands; however, a sequence-specific component depending on the  $+30$  core can easily be distinguished with the help of the mutant and wild-type competitor oligonucleotides (lanes 1–3). The probe  $-8/+112M$  (with 21 bp insertion at  $+15$ ) behaves very similarly to the probe  $-8/+34M$  in that it exhibits binding and competition properties analogous to those of the wild-type  $-8/+112$  probe, but again the intensity of the bands is decreased many-fold (lanes 4–6).

The DNA probes used reveal not only retarded bands with intermediate mobility, but usually tend to form a series of slowly migrating retarded bands as well, the whole picture being reminiscent of the assembly of polII transcriptional machinery on the promoter (Buratowski *et al.*, 1989). These series of bands presumably reflect the formation of multiprotein complexes on the promoter fragments, where not all of the bands exhibit sequence-specificity in binding. Most probably it is the identified protein that confers sequence-specificity to the entire transcriptional machinery in the case of the *mdg1* promoter. Identification of every band in the complex pattern requires purification of separate components and using them in reconstituted reactions.

The binding of the  $+30$  protein is not dependent on the RNA start-site sequence, since the efficiently competing oligonucleotide does not contain it. In addition, the  $+15/+112$  fragment was very similar to the  $-8/+112$  probe in its binding properties, i.e. it yielded an analogous pattern of bands when used as a probe and behaved identically when used as a competitor (data not shown). However, the difference in their transcriptional activity (Figures 2–4) indicates that, to enable efficient initiation of RNA synthesis, the  $+30$  core and the start-site sequence must act in combination. The most likely possibility is that the RNA start site is recognized by polII itself to initiate transcription at the appropriate distance from the binding site of the sequence-specific factor which directs efficient assembly of the transcription complex.

## Discussion

This study represents a detailed analysis of the promoter region of *D.melanogaster mdg1* retrotransposon, transcription of which is initiated precisely in the absence of TATA-box-like sequences in the appropriate position. The experiments presented here demonstrate that correct and precise initiation of *mdg1* RNA synthesis at basal levels is not dependent on the upstream non-transcribed sequences,

but can be provided only by the sequences extending from the transcriptional start site to the downstream  $+35$  region. The  $+30$  region is shown to be an efficient binding site for a sequence-specific protein factor, and mutation of the  $+30$  core sequences results in the absence of transcriptional activity in the *in vivo* assays.

To date, there are several known cases of localization of polII promoter sequences within the transcribed region near the RNA start site. The first is the *Drosophila* LINE element *jockey* (Mizrokhi *et al.*, 1988); transcription of *jockey* and several *mdg* elements is sensitive to  $\alpha$ -amanitin (Mizrokhi *et al.*, 1988). Two other *Drosophila* LINE elements, I and F (Fawcett *et al.*, 1986; DiNocera, 1988) were also inferred to possess internal promoters, because LINES cannot otherwise preserve their promoters after retroposition. The yeast Ty-D15 retrotransposon was reported to require an internal promoter element located 140 bp downstream of the LTR for its transcription (Yu and Elder, 1989).

Not only mobile elements but also some other *Drosophila* developmentally regulated cellular genes were shown to possess a similar promoter organization, i.e. correct initiation of their RNA synthesis at basal levels required not upstream, but downstream sequence elements (*Ubx*, *en*, *Antp*, E74; see above). It is of interest that nucleotide sequence comparisons allowed us to identify some similarities between them in the downstream region, and the identified core sequence was shown to be essential for *mdg1* transcription and for protein binding. Footprints observed for *Ubx* and *Antp* suggested the existence of the downstream protein factor(s); however, it was not investigated further with respect to its transcriptional importance and sequence-specificity of binding. Thus, the present work demonstrates for the first time the existence of a *D.melanogaster* sequence-specific DNA binding protein which is crucial for transcription initiation (we called it DIF, for downstream initiation factor). It remains to be seen what its relationship is to the other components of polII general transcriptional machinery and whether it represents a novel protein or an already known one but with an unknown function. The fact that the  $+30$  core bears no resemblance to the TATA box allows one to suggest that transcription of this type of promoter may not involve the only known sequence-specific general transcription factor TFIID, although it is possible that it may bind to non-TATA-like sequences. The exact sequence specificity of DIF binding has yet to be determined, as has its relationship to the other type of the  $+30$  elements.

As for mammalian and viral systems, two non-TATA and non-housekeeping GC-rich promoters have been examined in most detail: the lymphocyte-specific murine TdT (Smale and Baltimore, 1989; Smale *et al.*, 1990) and the SV40 major late promoter (Ayer and Dynan, 1988, 1990). The first one was shown to depend only on the RNA start-site sequence (called the Initiator or Inr), which in mammals looks quite similar to the previously known common *Drosophila* (and other insect) RNA start site  $TCA_{T}^{G}T_{C}^{T}$  (Arkipova *et al.*, 1986; Cherbas *et al.*, 1986; Hultmark *et al.*, 1986). However, one still cannot exclude the possibility of occasional occurrence of some sequence-specific protein-binding sites in the vicinity of Inr, since the Inr sequence does not seem to be the site of interaction for a sequence-specific general transcription factor. The  $TCA_{T}^{G}T_{C}^{T}$  sequence *per se* is necessary but not sufficient to promote accurate tran-



scription initiation, since the 3' promoter deletions of the +30 region for *Ubx*, *Antp*, *en*, *E74* and *mdg1* and mutation of the +30 core for *mdg1*, all of which retain the RNA start site and impair only the +30 region, demonstrate transcriptional inactivation of the previously active constructs.

On the other hand, the requirement of the RNA start-site sequence suggests a bipartite structure of a minimal promoter of this type, in which the two necessary elements carrying out the basal promoter function are the RNA start-site sequence, or *Inr*, and a downstream element at a proper distance. It may be that the initial event in transcription of such a promoter is binding of a sequence-specific component to the +30 core, followed by recognition of the TCAGT motif (maybe by polII itself), which is necessary for accurate initiation of RNA synthesis.

The SV40 major late promoter represents a somewhat more complicated tripartite structure in which the RNA start site as well as both regions around -30 and +30 are important for transcription, the start-site sequence directing precise initiation and the two others influencing its level (Ayer and Dynan, 1988). The +30 sequence is shown to be a site for interaction of a sequence-specific protein factor which facilitates formation of transcriptionally active preinitiation complexes (Ayer and Dynan, 1990). It is not yet clear whether transcription of this promoter is TFIID-dependent.

The mouse ribosomal protein L32 gene required an element ~30 bp downstream of the cap site for its transcription initiation, and a protein factor was shown to bind to it; however, the upstream sequences were required for accurate initiation of basal transcription as well (Moura-Neto *et al.*, 1989). TFIID was reported to provide efficient transcription from the adenovirus IVa2 promoter by binding to the TATA box in a downstream position (+20), and the *Inr* sequence was thought to fix the direction of transcription, since this downstream TATA box was present in an inverted orientation (Carcamo *et al.*, 1990). It was not clear whether TFIID was required for basal transcription from TdT promoter, although its regulated transcription was TFIID-dependent (Smale *et al.*, 1990).

In general, with respect to localization of sequences determining the transcription initiation, polII promoters may be divided into upstream promoters and downstream promoters. Upstream promoters may be defined as those which cannot be transcribed at all in the absence of upstream sequences; downstream ones are transcribed by the combination of sequences lying in close proximity to the RNA initiation site together with downstream elements, with upstream elements being needed only to regulate the level of transcription. The location of the RNA start site is likely to depend on the interaction of the start-site and downstream elements; it may be that both elements will reside in the transcribed region and then the promoter will be completely internal as for LINE elements.

It may turn out that the downstream promoter class is wider than initially apparent, because the regions responsible for initiation of RNA synthesis have not been mapped for many non-TATA promoters; furthermore, TATA box homologies are often ascribed in an arbitrary manner.

A brief look at the sequences of many other *Drosophila* genes (data not shown) which have a characteristic RNA start-site sequence, but in most cases do not have good

matches to the TATA box consensus, reveals the occurrence of the core sequence (CGTG or ACAG/C, or its reverse complement) at a distance of ~30–35 bp downstream of the transcriptional start site with a high degree of probability (much higher than the occurrence of any other four nucleotides) with a great variety of nucleotide sequences in between the two conserved areas. Of course, its significance should be investigated functionally in every particular case, since its occurrence may be occasional, and it will be of interest to find out whether the genes presented in Figure 5B do belong to the same promoter class as those demonstrated experimentally.

Identification of an internal polII promoter in LINE elements was predicted for a LINE-like I factor retrotransposon (Fawcett *et al.*, 1986), which does not generate terminal redundancy and needs another way to preserve its promoter after a retrotransposition cycle. The discovery of such a promoter organization for a retrovirus-like retrotransposon is rather intriguing, because *Drosophila* retrotransposons have been demonstrated to follow all the stages of retroviral reverse transcription leading to the formation of the LTR structure (Arkhipova *et al.*, 1986), and the LTRs are used for regeneration of the 5'-non-transcribed promoter regions during retrotransposition. In principle, it seems quite possible that such a promoter organization in some cases may result in a conversion of a retrovirus-like element into a LINE-like one with no need to regenerate promoter sequences for expression in an integrated site.

Thus, in addition to their evolutionary origin, the so-called viral and non-viral retrotransposons (Weiner *et al.*, 1986) may have much more in common than has been suggested, taking into consideration the possibility of interconversion between the two classes (it may theoretically occur in both directions as LINE-like elements may somehow acquire LTRs). Their relationship is strengthened by the existence of gag-like proteins in both classes (see Doolittle *et al.*, 1989 for references) and the presence of tRNA homologies and oligopurine stretches in the coding strands of several LINE elements (Murphy *et al.*, 1987; Furano *et al.*, 1988; Hutchison *et al.*, 1989). The possible significance and consequences of the existence of such promoters in regulated cellular and viral genes remain worthy of discussion and further investigation.

## Materials and methods

### Plasmid DNAs

The majority of the CAT constructs are based on the promoterless vector plasmid pUSVL-cat, which was created by insertion of a 1.6 kb *XmaI*–*BamHI* fragment from the plasmid pSVO-cat (Gorman *et al.*, 1982), which contained the coding sequences of the bacterial chloramphenicol acetyltransferase gene (with the ATG codon) followed by the SV40 intron/poly(A) site, into the *XmaI* and *EcoRI* sites of the pUC19 polylinker. This allowed further inserts to be cloned into the seven restriction sites from *HindIII* to *SmaI*.

*mdg1* fragments were usually excised from the plasmid p14 in which a 0.25 kb *EcoRI* LTR fragment (clone Dm58, Arkhipova *et al.*, 1986) is cloned into the *EcoRI* site of pUC19 so that the polylinker sites are adjacent to the 3' *EcoRI* LTR site in the transcribed region of *mdg1*. The promoter inserts of the 5'-deletion series -110/+112, -20/+112, -8/+112, +15/+112 were excised from p14 (by *HinPI*, *AccI*, *Sau3A* partial digest, and *BglII* respectively at the 5' LTR end and by *XmaI* in the polylinker at the 3' end) and inserted into pUSVL-cat digested by *XmaI* and by *SalGI*, *AccI*, *BamHI* and *BamHI* respectively. In the clone -95, the *TaqI*–*EcoRI*

fragment of p14 was inserted into the *AccI* and *SmaI* sites of pUSVL-cat; in the clone -31/+27, the *BglII*-*RsaI* fragment was inserted into the *BamHI* and *SmaI* sites. The insert of the entire LTR clone (catX) contained an extra 75 bp from the mdg1 sequences adjacent to the 3' LTR at the 5' end and 50 bp of the adjacent *D.melanogaster* genomic DNA sequences at the 3' end and was excised from the clone Dm58 by *DraI* and *XbaI* and inserted into the *SalGI* and *XbaI* sites of pUSVL-cat. The *BglII* fragment itself was also inserted into the *BamHI* site of pUSVL-cat in both orientations to obtain the constructs GD and GR. The construct -146 was obtained in a different way: the 1.6 kb *SmaI*-*BamHI* fragment of pSVO-cat was inserted into the plasmid p14 cut by these enzymes. It was further used to obtain the construct +27 by replacement of the *EcoRI*-*KpnI* fragment by the *RsaI*-*KpnI* fragment of p14.

To obtain the clones -8/+34 and -8/+34M, synthetic 21mers of the sequences GATCTGGAATGTACTCGCA (wild-type) and GATCTGGAATGGGATCCCGCA (mutant), respectively, were annealed to the complementary ones to produce a GATC overhang, phosphorylated and cloned into the plasmid -8/+112 cut by *BglII* and *SmaI* to replace the mdg1 sequence from +15 to +112 by the oligonucleotide. The clone -8/+112M was obtained by insertion of the mutant 21mer into the +15 *BglII* site of the same plasmid -8/+112. All the constructs were checked by dideoxy sequencing. Note that the *BglII* site at +15 is not present in a previously published mdg1 LTR sequence (Kulguskin *et al.*, 1982) due to a base error.

#### Transfection and CAT assay

Cell culture transfection by the standard calcium phosphate procedure and CAT assays with parallel  $\beta$ -galactosidase assays for normalization of CAT activity were performed as described by Mazo *et al.* (1989). The figures for CAT activity were usually calculated from three independent transfection experiments.

#### RNA extraction and analysis

Poly(A)<sup>+</sup> RNA was isolated from transfected cells and used for Northern blot hybridization and primer extension analysis according to Maniatis *et al.* (1982). In parallel with RNA extraction and analysis, CAT and  $\beta$ -galactosidase assays were performed on aliquots to control the RNA level and transfection efficiency. The <sup>32</sup>P-labelled (Multiprime DNA labelling system, Amersham) probe used for Northern experiments was the *HinfI* fragment of pUSVL-cat which contains the last 0.5 kb of the SV40 sequence, so only the transfected nucleic acids could be seen. The probe used for primer extension was synthesized on the M13cat single-stranded template described by Mazo *et al.* (1989).

#### DNA-protein interactions

Nuclear extracts were prepared from Schneider 2 tissue culture cells and assayed in gel retardation experiments essentially as described by Soeller *et al.* (1988). The binding reactions were performed on ice in 25 mM HEPES (pH 7.6), 40 mM KCl, 2 mM MgCl<sub>2</sub>, 0.1 mM EDTA, 1 mM DTT, 10% glycerol and 100  $\mu$ g/ml poly[dG-dC]·poly[dG-dC] for 20 min. In competition experiments, a 100-fold molar excess of cold DNA fragments isolated from polyacrylamide gels was added 10 min prior to addition of the labelled probe; oligonucleotides used for competition were phosphorylated, annealed, and ligated as described by Kadonaga and Tjian (1986). DNA-protein complexes were resolved in 4% (60:1) polyacrylamide gels in 0.25  $\times$  TBE at 20 V/cm.

#### Acknowledgements

The authors are indebted to L.Mizrokhi for providing the clone M13cat, to N.V.Lyubomirskaya for help, and to T.Paterson, C.McLean and especially D.Finnegan for support, providing unpublished data and critical reading of the manuscript. This work was supported in part by the Wellcome Trust fund in the UK.

#### References

Arhipova,I.R., Gorelova,T.V., Ilyin,Yu.V. and Schuppe,N.G. (1984) *Nucleic Acids Res.*, **12**, 7533-7548.  
 Arhipova,I.R., Mazo,A.M., Cherkasova,V.A., Gorelova,T.V., Schuppe,N.G. and Ilyin,Yu.V. (1986) *Cell*, **44**, 555-563.  
 Ayer,D.E. and Dynan,W.S. (1988) *Mol. Cell Biol.*, **8**, 2021-2033.  
 Ayer,D. and Dynan,W.S. (1990) *Mol. Cell Biol.*, **10**, 3635-3645.  
 Biggin,M.D. and Tjian,R. (1988) *Cell*, **53**, 699-711.  
 Bingham,P.M. and Zachar,Z. (1989) In Berg,D.E. and Howe,M.M. (eds),

*Mobile DNA*. American Society for Microbiology, Washington DC, pp. 495-502.  
 Boeke,J.D. (1988) In Domingo,E., Holland,J. and Ahlquist,P. (eds), *RNA Genetics*. CRC Press, Boca Raton, FL, Vol. 2, pp. 59-103.  
 Boeke,J.D. (1989) Berg,D.E. and Howe,M.M. (eds), *Mobile DNA*. American Society for Microbiology, Washington DC, pp. 335-374.  
 Boeke,J.D. and Corces,V.G. (1989) *Annu. Rev. Microbiol.*, **43**, 403-434.  
 Boeke,J.D., Garfinkel,D.J., Styles,C.A. and Fink,G.R. (1985) *Cell*, **40**, 491-500.  
 Bray,S.J. and Hirsh,J. (1986) *EMBO J.*, **5**, 2305-2311.  
 Buratowski,S., Hahn,S., Guarente,L. and Sharp,P.A. (1989) *Cell*, **56**, 549-561.  
 Carcamo,J., Maldonado,E., Cortes,P. Ahn,M.-H., Ilho-Ha, Kasai,Y., Flint,J. and Reinberg,D. (1990) *Genes Dev.*, **4**, 1611-1622.  
 Cherbas,L., Schultz,R.A., Koehler,M.M.D., Sarakis,C. and Cherbas,P. (1986) *J. Mol. Biol.*, **189**, 617-631.  
 DiNocera,P.P. (1988) *Nucleic Acids Res.*, **16**, 4041-4052.  
 DiNocera,P.P., Digan,M.E. and Dawid,I.B. (1983) *J. Mol. Biol.*, **168**, 715-727.  
 Doolittle,R.F., Feng,D.-F., Johnson,M.S. and McClure,M.A. (1989) *Quart. Rev. Biol.*, **64**, 1-30.  
 Driver,A., Lacey,S.F., Cullingford,T.E., Mitcheson,A. and O'Hare,K. (1989) *Mol. Gen. Genet.*, **220**, 49-52.  
 Dynan,W.S. (1986) *Trends Genet.*, **2**, 196-197.  
 Dynan,W.S. (1989) *Cell*, **58**, 1-4.  
 Emori,Y., Shiba,T., Kanaya,S., Inouye,S., Yuki,S. and Saigo,K. (1985) *Nature*, **315**, 773-776.  
 Fawcett,D.H., Lister,C.K., Kellett,E. and Finnegan,D.J. (1986) *Cell*, **47**, 1007-1015.  
 Finnegan,D.J. and Fawcett,D.H. (1986) *Oxford Surveys on Eukaryotic Genes*, **3**, 1-62.  
 Friesen,P., Rice,W.C., Miller,D.W. and Miller,L.K. (1986) *Mol. Cell Biol.*, **6**, 1599-1607.  
 Furano,A.V., Robb,S.M. and Robb,F.T. (1988) *Nucleic Acids Res.*, **16**, 9215-9231.  
 Georgiev,G.P. (1984) *Eur. J. Biochem.*, **145**, 203-220.  
 Gorman,C.M., Moffat,L.F. and Howard,B.H. (1982) *Mol. Cell Biol.*, **2**, 1044-1051.  
 Hirsh,J., Morgan,B.A. and Scholnick,S.B. (1986) *Mol. Cell Biol.*, **6**, 4548-4557.  
 Hultmark,D., Klemenz,R. and Gehring,W.J. (1986) *Cell*, **44**, 429-438.  
 Hutchison,C.A., III, Hardies,S.C., Loeb,D.D., Shehee,W.R. and Edgell,M.H. (1989) In Berg,D.E. and Howe,M.M. (eds), *Mobile DNA*. American Society for Microbiology, Washington DC, pp. 593-618.  
 Ilyin,Yu.V., Chmeliauskaitė,V.G. and Georgiev,G.P. (1980) *Nucleic Acids Res.*, **8**, 3439-3457.  
 Inouye,S., Hattori,K., Yuki,S. and Saigo,K. (1986) *Nucleic Acids Res.*, **14**, 4765-4778.  
 Kadonaga,J.T. and Tjian,R. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 5889-5893.  
 Kassis,J.A., Desplan,C., Wright,D.K. and O'Farrell,P.H. (1989) *Mol. Cell Biol.*, **9**, 4304-4311.  
 Kulguskin,V.V., Ilyin,Y.V. and Georgiev,G.P. (1982) *Genetika*, **18**, 869-879.  
 Lewin,B. (1990) *Cell*, **61**, 1161-1164.  
 Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.  
 Mathis,D.J. and Chambon,P. (1981) *Nature*, **290**, 310-315.  
 Matsuo,Y., Kugimiya,W., Kadokami,Y. and Saigo,K. (1986) *Nucleic Acids Res.*, **14**, 9521-9622.  
 Mazo,A.M., Mizrokhi,L.J., Karavanov,A.A., Sedkov,Y.A., Krichevskaja,A.A. and Ilyin,Yu.V. (1989) *EMBO J.*, **8**, 903-911.  
 Mizrokhi,L.J., Georgieva,S.G. and Ilyin,Yu.V. (1988) *Cell*, **54**, 685-691.  
 Moura-Neto,R., Dudov,K.P. and Perry,R.P. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 3997-4001.  
 Murphy,N.B., Pays,A., Tebabi,P., Coquelet,H., Steinert,M. and Pays,E. (1988) *J. Mol. Biol.*, **195**, 855-871.  
 O'Hare,K., Murphy,C., Levis,R. and Rubin,G.M. (1984) *J. Mol. Biol.*, **180**, 437-455.  
 Perkins,K.K., Dailey,G.M. and Tjian,R. (1988) *Genes Dev.*, **2**, 1615-1626.  
 Pirrotta,V., Manet,E., Hardon,E., Bickel,S.E. and Benson,M. (1987) *EMBO J.*, **6**, 791-799.  
 Priimagi,A.F., Mizrokhi,L.J. and Ilyin,Yu.V. (1988) *Gene*, **70**, 253-262.  
 Regulski,M., McGinnis,R., Chadwick,R. and McGinnis,W. (1987) *EMBO J.*, **6**, 767-777.



- Shenk, T. (1981) *Curr. Top. Microbiol. Immunol.*, **93**, 25–46.
- Smale, S.T. and Baltimore, D. (1989) *Cell*, **57**, 103–113.
- Smale, S.T., Schmidt, M.C., Berk, A.J. and Baltimore, D. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 4509–4513.
- Sneddon, A. and Flavell, A.J. (1989) *Nucleic Acids Res.*, **17**, 4025–4035.
- Soeller, W.C., Poole, S.J. and Kornberg, T. (1988) *Genes Dev.*, **2**, 68–81.
- Spana, C., Harrison, D.A. and Corces, V.G. (1988) *Genes Dev.*, **2**, 1414–1423.
- Struhl, K. (1987) *Cell*, **49**, 295–297.
- Thummel, C.S. (1989) *Genes Dev.*, **3**, 782–792.
- Wasylyk, B. (1988) *Biochim. Biophys. Acta*, **951**, 17–35.
- Weiner, A.M., Deininger, P. and Efstratiadis, A. (1986) *Annu. Rev. Biochem.*, **55**, 631–661.
- Wilde, C.D. and Akam, M. (1987) *EMBO J.*, **6**, 1393–1401.
- Xiong, Y. and Eickbush, T. (1988) *Mol. Cell. Biol.*, **8**, 114–123.
- Yu, K. and Elder, R.T. (1989) *Mol. Cell. Biol.*, **9**, 3667–3678.
- Yuki, S., Inouye, S., Ishimaru, S. and Saigo, K. (1986) *Eur. J. Biochem.*, **158**, 403–410.

Received on July 12, 1990; revised on January 16, 1991