The EMBO Journal vol.6 no.9 pp.2773-2779, 1987

Coincident start sites for divergent transcripts at a randomly selected CpG-rich island of mouse

Patrizia Lavia[1], Donald Macleod and Adrian Bird[2]

MRC Mammalian Genome Unit, King's Buildings, West Mains Road, Edinburgh EH9 3JT, Scotland, UK

[1]Present address: Centro di Genetica Evoluzionistica CNR, c/o Dipartimento di Genetica e Biologia Molecolare, I Universita di Roma, Rome 00185, Italy

[2]Present address: MRC Clinical and Population Cytogenetics Unit, Western General Hospital, Crewe Road, Edinburgh, Scotland, UK

Communicated by A.Bird

We determined the nucleotide sequence of two HTF islands that were selected at random from mouse chromosomal DNA. Both were non-methylated, G+C rich, and contained CpG at close to the expected frequency. When used as probes, the two islands detected multiple transcripts in RNA from several mouse tissues. Cloned cDNAs for the major transcripts of one island (HTF9) were isolated and used to construct a transcriptional map. We found that HTF9 contains the origin of a pair of divergent transcripts that are probably messenger RNAs. The bidirectional promoter is different from those previously observed as the major transcription start sites for each orientation are coincident on opposite strands of the DNA. The results support the view that HTF islands often mark genes, and they suggest that bidirectional transcription may be a common feature of island promoters.

Key words: DNA methylation/HTF islands/bi-directional transcription

Introduction

The chromosomal DNA of vertebrates is on average A+T rich and contains the sequence CpG at about one-fifth of the frequency predicted from base composition. CpG is often methylated at the 5 position of cytosine, and as a result restriction endonucleases that are prevented from cutting by CpG methylation (for example *Hpa*II) cleave vertebrate DNA poorly. A small fraction of the genome, however, is extensively cleaved by *Hpa*II (Cooper *et al.*, 1983), and hence has been called the *Hpa*II tiny fragment (HTF) fraction. We previously cloned a selection of HTFs from mouse liver DNA and characterized three in detail (Bird *et al.*, 1985). All three belong to 'islands' of genomic DNA that are neither methylated nor CpG deficient. Sequences with these general properties have been found surrounding the 5′ ends of 'housekeeping' genes and several tissue specific genes (McKeon *et al.*, 1982; Stein *et al.*, 1983; Tykocinski and Max, 1984; Bird, 1986).

The HTF fraction could accommodate ~30 000 HTF islands per haploid genome. An attractive possibility is that most of these are derived from the 5′ domains of genes. Another possibility, however, is that most HTF islands are not associated with genes; the tendency to study genes rather than intergenic DNA may have led to the preferential identification of gene-associated islands. In order to investigate these possibilities we have asked if the three random islands that were originally isolated from mouse DNA are associated with transcripts. Our results sustain the view that most HTF islands are associated with genes.

Results

The sequence of HTF9 and HTF12

The identification and partial characterization of three HTF islands from mouse genomic DNA (HTFs 9, 12 and 5) has been described (Bird *et al.*, 1985). The islands were isolated by probing a genomic library with random cloned fragments from the HTF fraction. Figure 1 shows the distribution of CpGs in two of the islands (HTFs 9 and 12) whose sequence we have determined. The sequence of HTF9 and its flanking DNA is shown in Figure 2. Within the CpG clusters all testable sites are nonmethylated in DNA from several mouse tissues (Figure 1A). In both cases the island region is G+C rich whereas flanking regions have a G+C content that is typical of bulk DNA (Figure 1B). The density of CpGs within both islands is close to the expected density for DNA of this base composition, although there are marked downward fluctuations within HTF9 (Figure 1C). Outside the islands CpG density is considerably lower than that of GpC, as expected for bulk DNA.

Transcripts

*Eco*RI fragments containing HTFs 9, 12 and 5 were used to probe Northern blots of polyadenylated RNA from several mouse tissues. Probes 9 (pL9.2) and 12 (pL12.15) are single copy as judged by Southern blots, while probe 5 shows some cross hybridization to moderately repeated sequences (Bird *et al.*, 1985). HTF9 detected two groups of transcripts at ~2.7 and 0.7-1.0 kb in all seven mouse tissues that were tested (Figure 3, lanes a-g) and in cultured L cells (not shown). HTF12 detected a pair of transcripts in liver, brain (Figure 3, lanes h and i), kidney, testis and L-cell RNA (not shown). Other tissues were not tested with this probe. We noted that the lower band of the doublet coincided with the large ribosomal RNA, and might therefore represent spurious homology with this sequence. The signal in this band, however, was not affected by two cycles of purification of the RNA by oligo-dT cellulose, or by the addition of excess rRNA to the hybridization mix (not shown). This implies that both bands represent polyadenylated transcripts complementary to HTF12. Probe HTF5 gave no reproducible signal in liver, kidney, testis or brain RNA under conditions where the repeated sequences present in the probe were fully competed (not shown).

Two transcripts from HTF9

We studied island HTF 9 in more detail in order to determine its location with respect to the transcripts. Initially, Northern blots of poly(A)$^+$ RNA from mouse liver were probed with genomic sequences flanking pL9.2 to the left (pL9.3) and right (pL9.5). Surprisingly, pL9.3 hybridized to only the 0.7-1.0-kb RNAs, while pL9.5 hybridized only to the 2.7-kb band (Figure 4, lanes a-c). We next restricted pL9.2 so as to generate probes including different portions of HTF9 itself (probes A and C, Figure 4). A Northern blot hybridization experiment using these probes confirmed that the small RNA cluster could only be detected by the

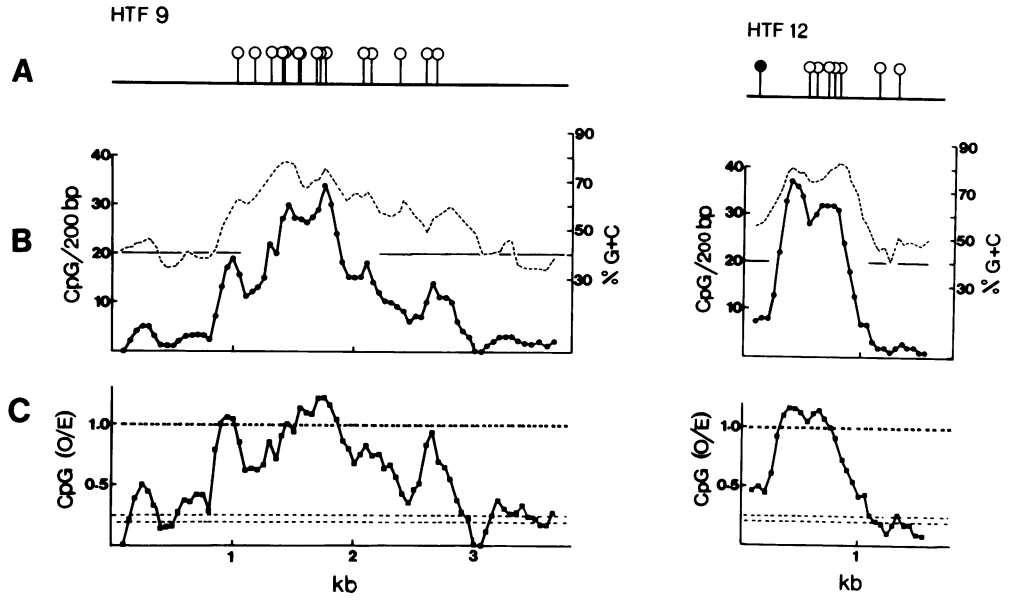

Fig. 1. Lack of methylation and the frequency of CpG at HTF islands 9 and 12 of mouse. (**A**) Cloned *Eco*RI fragments pL9.2 and pL12.15 which contain islands 9 and 12 respectively showing CpG enzyme sites that have been tested for methylation in previous studies (Bird *et al.*, 1985; Brown and Bird, 1986). Open circles, unmethylated sites; solid circle, methylated site. (**B**) Graph of CpG frequency (solid line, left ordinate) and percentage G+C content (broken line, right ordinate) per 200-bp segment (step of 50 bp) through the *Eco*RI fragment above. The straight line at 40% G+C shows the average base composition of mouse chromosomal DNA. (**C**) Observed over expected (O/E) frequency of CpG per 200 bp (step of 50 bp). The single broken line shows an O/E ratio of 1.0. The double broken line gives the average value of O/E for bulk genomic DNA (0.2−0.25).

genomic region left of HTF9, whereas the large transcript was detected by sequences to the right (Figure 4, lanes d and e). We designated the small and large transcripts as RNA-A and RNA-C respectively.

Formally the results shown in Figure 4 could be interpreted in two ways. Either left and right probes were detecting different splicing products from a single transcriptional unit, or these are two different transcription units which diverge, converge or overlap at HTF9. Since it is known for many genes that HTF sequences lie at the 5′ end of transcription units, we were interested in the possibility that HTF9 was at the 5′ end of RNAs A and C; in other words that the A and C RNAs are transcribed divergently, on opposite strands, from HTF9. In order to test this, the *Hin*dIII central fragment of pL9.2, which includes the entire HTF9 island (see Figure 4), was subcloned into M13 in both orientations with respect to the primer annealing site, and the subclones were used to generate uniformly labelled, opposite-strand probes. When these were hybridized to poly(A)$^+$ RNA, both strands of pL9.2 were found to be transcribed (Figure 4, lanes f and g). The (+) strand hybridized to RNA-A, while the (−) strand hybridized to RNA-C (Figure 4, lanes f and g). The direction of the primed single-strand synthesis showed that the A RNAs are transcribed leftwards from pL9.2, while the C RNA is transcribed rightwards. Thus HTF9 is located at the 5′ end of two opposite-strand, divergent transcripts. It is of interest that when the same experiment was performed with HTF12, both transcript bands annealed to the same strand of the probe. Thus there is no indication of bi-directional transcription at this island.

Isolation and characterization of cDNA clones

To further characterize the divergent transcripts, a mouse embryo cDNA library cloned in λ gt10 was screened with probes A and C (see Figure 4). Screening of 5 × 10^5 clones with probe A yielded 11 positives ranging from 0.7 to 1 kb in length. Screening with probe C yielded one positive of 2.3 kb in length. The sizes of A cDNA clones agreed with those obtained from the Northern blots, while the C cDNA clone was ~400 bp shorter than the observed transcript. The 10-fold difference in the abundance of the A and C cDNA clones in the library may reflect different transcript abundance, or under-representation of the longer cDNAs in the library.

The cDNA clone selected with probe C contained three *Eco*RI fragments, which were ordered by partial *Eco*RI digestion and hybridization to the HTF9-containing *Hin*dIII fragment from pL9.2. The fragment hybridizing to the island was then subcloned into M13 and sequenced from both ends. Projection of the cDNA sequence onto the genomic sequence showed three regions of homology with the genomic sequence of pL9.2 (Figure 7). Each interruption of the homology with the genome showed a typical exon/intron junction consensus sequence (Sharp, 1981; Mount, 1982; see Figure 2) indicating that the primary transcript is spliced *in vivo*. We have no direct evidence for translation of this transcript. However, the three mapped exons, representing 41% of the entire cDNA length, show a single open reading frame (ORF) starting at the ATG located at position 2161 in pL9.2 and extending for 894 bp to the *Eco*RI site which marks the 3′ boundary of our sequence. The predicted amino acid sequence encoded by this 5′ portion of RNA C is shown in Figure 5.

The cDNA clones selected with probe A were analysed in a similar way. Restriction mapping of the cDNAs showed a major *Eco*RI fragment of 720−810 bp in different clones. This fragment hybridized to pL9.2 and was therefore likely to contain the 5′ region of the cDNAs. In most cDNA clones we detected one more *Eco*RI fragment, whose length varied between 40 and 330 bp. This finding indicated that the size heterogeneity observed within the A RNA cluster in the Northern blots was primarily due to variation near the 3′ end. The major *Eco*RI fragment was subcloned from three independent cDNA clones and was sequenced from both ends. The regions of homology with the genome were then mapped. Each of them was found to be interrupted by typical splice sites in the genome (see Figure 2). The first region of homology was found to be entirely contained within

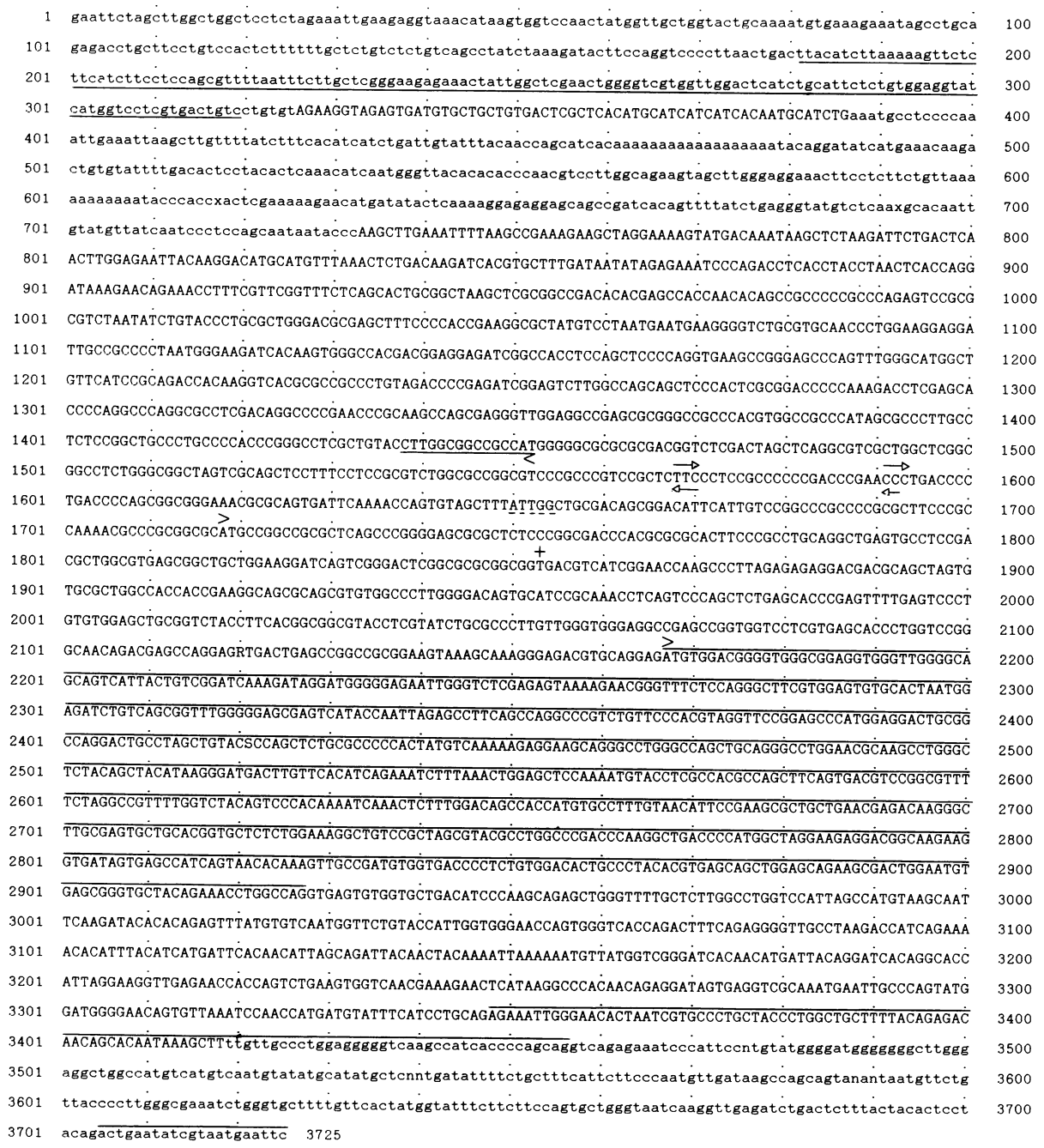

Fig. 2. The sequence of HTF9 and flanking sequences. A single strand corresponding to the insert of plasmid pL9.2 is shown (see Figure 1 and Bird *et al.*, 1985). Capitals indicate sequences that were determined on both DNA strands, and lower-case letters show regions where only one strand was sequenced. Unbroken underlines and overlines are potential coding exons for the opposite strand transcripts RNA-A and RNA-C respectively (see Figure 7 for a diagrammatic representation). Arrows mark major transcription start sites in each direction. The broken underline is a potential CCAAT box in the same orientation as transcript A. Potential translational initiation codons (AUG) are marked by open arrowheads. One of the three AUGs is at the beginning of an ORF in the proposed leader sequence of transcript C. Its termination is marked by +.

the island region and was spliced onto a second exon also within the *Eco*RI fragment (see Figure 7). The genomic location of exons further downstream is unknown. When the 5′ region obtained from three overlapping cDNAs was analysed, a single ORF of 228 bp was detected, preceded by a leader region of ~90 bp in the largest cDNA. The sequence at the 3′ end of this *Eco*RI fragment was identical in all three cDNA clones. At this end, stop codons were present in all three frames, allowing us to estimate that the coding region of gene A must be <460 bp long. The predicted amino acid sequence for A (Figure 5) represents the first 76 residues from the translation start codon, which according to our estimate accounts for about half of the expected A protein.

Mapping of 5′ ends

Transcription start sites for both A and C RNAs were determined by S1-nuclease mapping (Figure 6). For A the probe was a single-stranded fragment of genomic DNA extending from a labelled *Taq*I site 20 bp upstream of the putative AUG (position 1466) to a *Sma*I site 265 bp further upstream (position 1731).

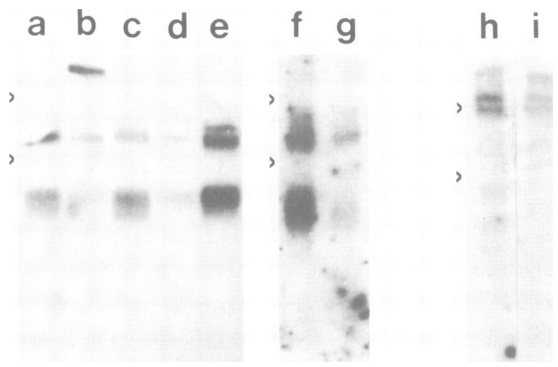

Fig. 3. Discrete transcripts in RNA from several mouse tissues detected by pL9.2 and pL12.15. **Lanes a–g**, poly(A)$^+$ RNA from intestine (**a**), submaxillary gland (**b**), spleen (**c**), kidney (**d**), testis (**e**), liver (**f**) and brain (**g**) probed with pL9.2. **Lanes h** and **i**, poly(A)$^+$ RNA from liver (**h**) and brain (**i**) probed with pL12.15. The upper and lower arrow heads show the positions on these gels of 28S and 18S ribosomal RNAs respectively.

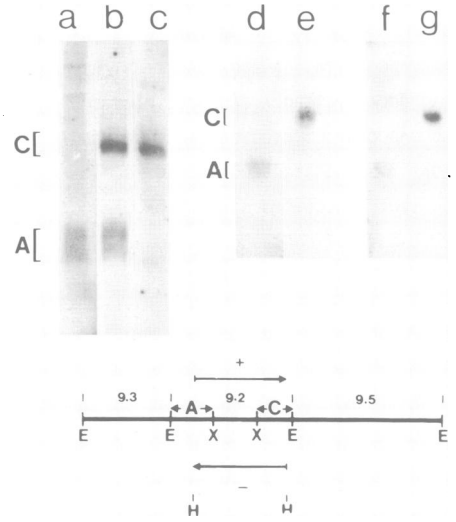

Fig. 4. Evidence for bi-directional transcription from island HTF9. Blots of poly(A)$^+$ RNA from liver were hybridized to the probes diagrammed below the figure. **Lanes a–c** show bands detected by the contiguous genomic subclones pL9.3 (**a**), pL9.2 (**b**) and pL9.5 (**c**). **Lanes d** and **e** show bands detected by subfragments A (**d**) and C (**e**) of pL9.2. Probes A and C are each bounded by an *Xho*I site (X) and an *Eco*RI site (E) as indicated. **Lanes f** and **g** show the results of hybridizing opposite strands of pL9.2 to poly(A)$^+$ RNA. Arrows (+) and (–) show the length and direction of synthesis of the probes, which are derived from an internal *Hin*dIII (H) fragment of pL9.2 subcloned into an M13 vector.

After annealing with total liver RNA and trimming with S1 nuclease, two major and several minor protected fragments were visible on a sequencing gel. Variation of the temperature and enzyme concentration during nuclease treatment affected the overall intensity of the bands, but not their intensities relative to one another (Figure 6). This indicated that multiple bands were not the result of S1 nicking within the RNA–DNA hybrid. Runs of A or T, which can give rise to such artefacts, are notably absent from this region of the sequence. The unlikely possibility that the multiple bands are due to several alternative splices onto an exon that lies further upstream is rendered highly improbable by the absence of consensus splice sites in this region of genomic DNA. We conclude that, in common with several housekeeping genes, RNA-A is initiated at multiple sites. About

'A'

```
  1  MetAlaAlaAlaArgThrValHisGluAspHisAspThrSerThrGluAsnAlaAspGlu    60
 61  SerAsnHisAlaProGlnPheGluProIleValSerValProGluGlnGluIleLysThr   120
121  LeuGluGluAspGluGluGluLeuPheLysMetArgAlaLysLeuPheProValCysPhe   180
181  ArgGluAsnAspLeuProGluTrpGluGlyAlaArgHis  219
```

'C'

```
  1  MetTrpThrGlyTrpAlaGluValGlyTrpGlySerSerHisTyrCysArgIleLysAsp    60
 61  ArgMetGlyGluAsnTrpValSerArgValLysGluArgValSerProGlyLeuArgGly   120
121  ValCysThrAsnGlyAspLeuSerAlaValTrpGlySerGluSerTyrGlnLeuGluPro   180
181  SerAlaArgProValCysSerHisValGlySerGlyAlaHisGlyGlyLeuArgProGly   240
241  LeuProSerCysThrProAlaLeuArgProHisTyrValLysLysArgLysGlnGlyLeu   300
301  GlyGlnLeuGlnGlyLeuGluArgLysProGlyLeuTyrSerTyrIleArgAspAspLeu   360
361  PheThrSerGluIlePheLysLeuGluLeuGlnAsnValProArgHisAlaSerPheSer   420
421  AspValArgArgPheLeuGlyArgPheGlyLeuGlnSerHisLysIleLysLeuPheGly   480
481  GlnProProCysAlaPheValThrPheArgSerAlaAlaGluArgAspLysGlyLeuArg   540
541  ValLeuHisGlyAlaLeuTrpLysGlyCysProLeuAlaTyrAlaTrpProAspProArg   600
601  LeuThrProTrpLeuGlyArgGlyArgGlnGluGlyAspSerGluProSerValThrGln   660
661  SerCysArgCysGlyAspProSerValAspThrAlaLeuHisValSerSerTrpSerArg   720
721  SerAspTrpAsnValSerGlyCysTyrArgAsnLeuAlaArgGluIleGlyAsnThrAsn   780
781  ArgAlaLeuLeuProTrpLeuLeuLeuGlnArgGlnGlnHisAsnLysAlaPheValAla   840
841  LeuGluGlyValLysProSerProGlnGlnThrGluTyrArgAsnGluPhe  891
```

Fig. 5. The predicted amino acid sequences encoded by the 5′ portions of transcripts A and C. The predictions were deduced from sequence data for the shaded regions of the cDNA clones shown in Figure 7(B). See also under- and over-lined regions of the sequence in Figure 2.

55 bp upstream of the major start site is the sequence CCAATAAA which contains a perfect CCAAT homology, but also an ATAAA (TATA-like) homology.

The cDNA clone for RNA-C is ~500 bp shorter than the length of the transcript observed on Northern blots. If no introns lay between the 5′ end of the cDNA clones and the 5′ end of the transcript, then we expected that an S1 probe encompassing the region 500 bp upstream of the cDNA would detect transcription start sites. Using a *TaqI–XmaI* probe (Figure 6) we observed major protected fragments of ~147 and 171 bp plus additional minor bands indicating that RNA-C, like RNA-A, has multiple transcription start sites. Surprisingly, the two major 5′ ends for RNA-C coincide within a few nucleotides with the major 5′ ends of RNA-A. The relatively large distance between the putative AUG for transcript C and the transcription start sites led us to check that short introns in this region had not been overlooked. A *PstI–BglII* fragment (Figure 6) covering this region was used as a probe in S1 experiments. No bands indicative of splice sites were seen, and the fragment appeared fully protected by RNA (not shown). Thus the first exon of RNA-C is >1.3 kb in length. If the AUG at the beginning of the long ORF (position 2161) is indeed the initiator codon, then the untranslated leader for this putative mRNA is ~570 bp long. A short ORF of 135 bp is present in the leader (Figure 2). Short ORFs have been seen in the leader sequences of several vertebrate and yeast mRNAs. Their possible significance has been discussed by Kozak (1986). Figure 7(A and B) summarizes the mapping data relating HTF9 to its transcripts and Figure 7C shows a detailed map of major and minor transcription start sites.

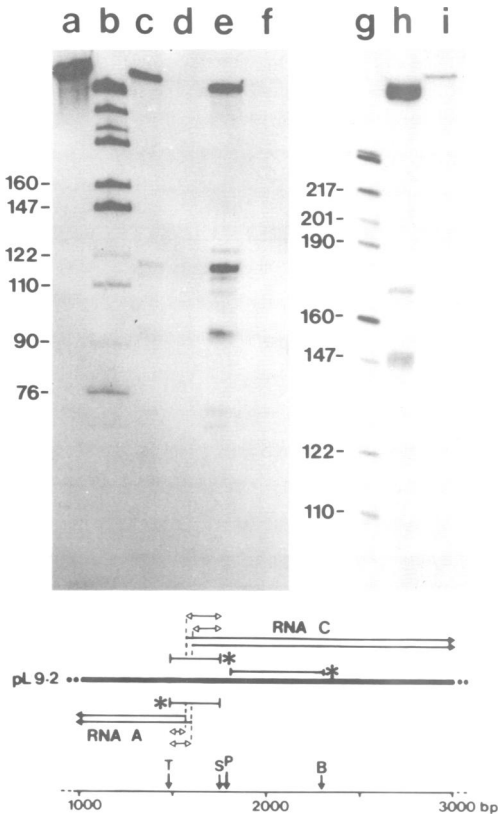

Fig. 6. Location of the 5′ ends of transcripts A and C by S1-nuclease analysis. The probes were opposite strands of a *Taq*I (T)−*Xma*I (S) fragment labelled at their 5′ ends (asterisks) as shown in the diagram. Labelled probes were annealed with total RNA of mouse liver or yeast transfer RNA as described in Materials and methods. **Lanes a−f**, S1 analysis of transcript A showing the full sized probe (**a**), a marker lane of plasmid pAT153 digested with *Hpa*II (**b**), results of treating the mouse liver RNA−DNA hybrids with two concentrations of S1 (**c** and **e**), (e) being 3-fold less than (c). We strongly suspect that this batch of S1 nuclease was less active than predicted on the label. The number of units of enzyme used in these experiments is therefore not known. **Lanes d** and **f**, the same two S1 concentrations as (c) and (e), but after annealing of the probe to 20 μg yeast transfer RNA. **Lanes g−i**, analysis of transcript C showing marker as lane b (**g**), probe annealed to 50 μg total liver RNA (**h**), probe annealed to 20 μg yeast transfer RNA (**i**). The diagram also shows a probe bounded by *Pst*I (P) and *Bgl*II (B) sites labelled at the *Bgl*II end. This probe was used to scan for start sites or splice sites further downstream. None was found (see text, data not shown). Coordinates on the diagram refer to the sequence of pL9.2 as shown in Figure 2.

Discussion

HTF islands often mark genes

In vertrebrates, all known polymerase II 'housekeeping' genes, and a proportion of known tissue-specific genes, have sequences with the properties of HTF islands surrounding their 5′ ends (reviewed in Bird, 1986). The islands are usually 1−2 kb in length, and include the first one or two exons of the gene, as well as upstream sequences. These findings, together with the similarity between the approximate number of HTF islands per genome (30 000) and the anticipated number of genes (~50 000), gave rise to the hypothesis that the majority of islands are located at the 5′ ends of genes. In support of the hypothesis, we found in this study that two out of three randomly selected islands detect transcripts on Northern blots, and that the one island which was characterized further has a 5′ location with respect to two RNAs. The absence of a transcript corresponding to HTF5 in four mouse tissues shows that this island may not be associated with any gene. It cannot be excluded, however, that HTF5 is part of a tissue-specific gene whose transcript is not present in the tissues that were tested. HTF islands have been observed at several tissue-specifically expressed genes; for example chicken α2(I) collagen (McKeon *et al.*, 1982), mouse Thy-I (Kolsto *et al.*, 1986), and human α globin (Bird *et al.*, 1987).

Two unrelated RNAs are transcribed divergently from coincident sites near the centre of the island HTF9. Both transcripts are spliced, polyadenylated, and are found in all seven mouse tissues that were tested, as well as cultured L cells. Both transcripts are also present early in development, as the source of the cDNA clones was an 8.5-day mouse embryo cDNA library. We do not have direct evidence that A and C correspond to messenger RNAs encoding proteins, but indirect evidence favours this view. In particular the presence in the cDNAs of long ORFs following a potential AUG initiator codon point to a protein-coding function. Also, RNAs of similar size to A and C can be detected by the island probe in polyadenylated RNA from rat liver and HeLa cells (J.Lewis, unpublished observations). The evolutionary conservation of RNA sequences would not be expected unless there is stringent selection pressure to conserve these sequences. Taken together, the evidence suggests that A and C transcripts code for two proteins with a 'housekeeping' function in the cell. Since the polyadenylated transcripts detected by HTF12 were also observed in all tested tissues, there is a possibility that these RNAs too encode housekeeping proteins. The frequent presence of HTF islands at the promoters of housekeeping genes may tell us something about their function. Housekeeping promoters must be constantly available to nuclear transcription factors, and it is possible that the unusual sequence composition and lack of methylation of island DNA insures this. For example, island sequences may influence chromatin structure, either directly or via bound factors (Bird, 1986).

HTF9 contains a bidirectional promoter

The 5′ ends of transcripts A and C map to a pair of coincident sites in HTF9 located close to the most CpG-rich/G+C-rich region of the island (cf. Figures 1 and 7). Processing of the 5′ ends of RNAs transcribed by polymerase II has not been observed, and it is widely believed that the 5′ ends of cytoplasmic transcripts represent the first nucleotides that were transcribed in the nucleus. In this case, the HTF9 promoter is genuinely bidirectional, as polymerases can transcribe to the right or to the left from the same sequence of <5 bp. Some features of the genomic sequence in this region are noteworthy. The region contains two copies of the consensus binding sequence for the protein Sp1 (Figure 7C). In addition to its importance for viral transcription, this protein has been implicated in promoter function for several cellular genes, all of which have HTF-island-like sequences surrounding their promoters (Kadonaga *et al.*, 1986; Dynan, 1986). It may be significant that, in spite of its asymmetrical binding sequence, Sp1 can activate transcription in either orientation. One of the GC boxes in HTF9 has the sequence of a strong binding site for Sp1, and is located between the two transcriptional start sites. A related sequence is present close to the initiation sites for the mouse HPRT gene, and has been shown to be a crucial element of the HPRT promoter (Melton *et al.*, 1986).

There is no TATAA homology upstream of transcript C, but a related sequence does exist as CCAATAAA >50 bp upstream of A (Figure 7C). Coincidentally, this sequence also contains the CCAAT homology, which is not found elsewhere near the

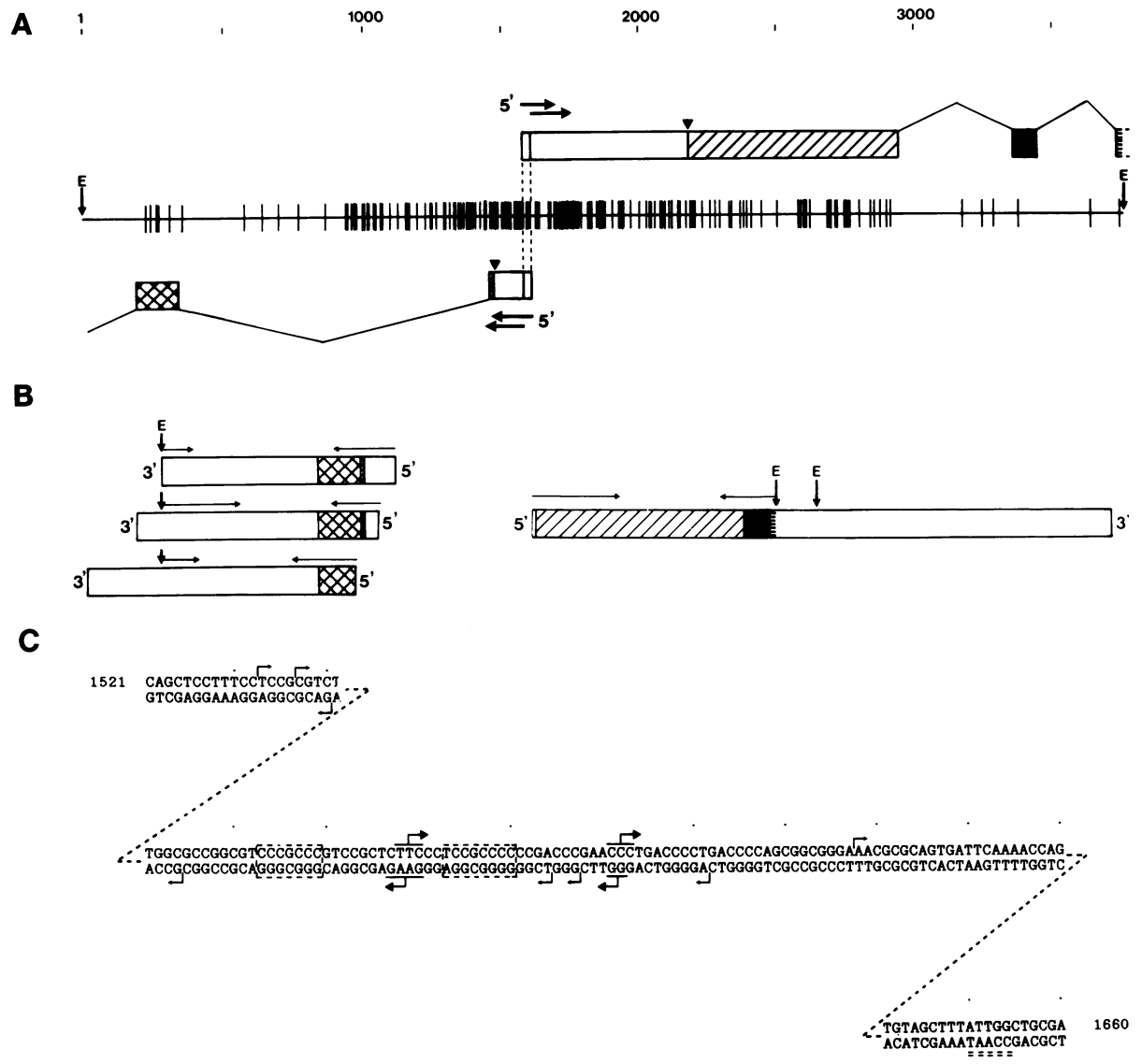

Fig. 7. Divergent transcription units initiating at coincident sites in HTF island 9. (**A**) The 3725-bp *Eco*RI (E) fragment that was subcloned in pL9.2 is shown with CpGs marked by vertical lines. The region of high CpG density from 1000 bp to 3000 bp corresponds to the HTF island (see Figure 1). The 5′ portion of divergent transcripts C (above the map) and A (below the map) are shown. Exons are shaded to allow comparison with the cDNA maps below. Horizontal arrows marks the two major transcriptional start sites for A and C. Dotted lines emphasize their coincidence. Solid arrow heads show putative translation initiator codons that are followed by long ORFs in the cDNA sequences. (**B**) Diagrams of three characterized cDNAs for transcripts A (left) and one for transcript C (right). The two classes of cDNA are drawn in opposite orientations to permit comparison with the divergent transcription diagram shown above. Vertical arrows show *Eco*RI sites (E). Horizontal arrows cover regions whose nucleotide sequence was determined. The scale is as above. (**C**) Detailed sequence map of the region containing transcription start sites for A and C. Large arrows denote major start sites, and small arrows minor starts. Broken line boxes contain potential Sp1 binding sites. A potential CCAAT box is underlined (double broken line). Coordinates refer to the sequence in Figure 2.

promoter. In particular there is no CCAAT homology upstream of the transcription start sites for C. The abnormally large distance between the TATAA homology and the transcription start site for A, and the presence of multiple 5′ ends for this transcript, argue against a TATAA-box function for this sequence. There are now several examples of promoters without TATAA boxes (e.g., Reynolds *et al.*, 1984; Melton *et al.*, 1984; Crouse *et al.*, 1985; Giguere *et al.*, 1985), many of which have been shown to give rise to transcripts with heterogeneous 5′ ends. Without the discipline imposed by a TATAA box, it is possible that initiation of transcription within accessible regions of the genome is relatively relaxed with respect to position and orientation.

There is evidence that divergent transcription may occur frequently at HTF islands. Opposite strand transcripts that initiate in the vicinity of the DHFR major transcription start sites have been detected in mouse (Farnham *et al.*, 1985; Crouse *et al.*, 1985) and Chinese hamster ovary cells (Mitchell *et al.*, 1986) that have amplified the DHFR gene. The status of the divergent transcript reportedly differs between cell types, being a small, nuclear, non-polyadenylated RNA in one case, and a relatively large, cytoplasmic, polyadenylated RNA in the other case (discused by Mitchell *et al.*, 1986). Earlier studies suggested that a segment of African green monkey DNA can promote transcription in both directions (Saffer and Singer, 1984). The sequence, which contains homology to the 21-bp repeat of SV40, gave transcription regardless of its orientation in a eukaryotic expression vector, and there was evidence for initiation of transcription in one, and perhaps two, directions *in vivo*. Recently another pair of divergent transcripts has been observed in the mouse (Williams and Fried, 1986). Opposite transcription start sites occur

50–75 bp apart within a sequence which, on the basis of its CpG content, resembles an HTF island. Coincident initiation sites for bi-directional transcription as identified within HTF9 have not previously been reported. It seems likely that in this case promoter elements are shared by transcripts A and C, and therefore that mutation of elements that affect the efficiency of A transcription will affect C transcription in a similar way. If so, an intriguing consequence is that elements that are located upstream of one transcription start site are downstream of another.

Materials and methods

Sequencing

The nucleotide sequence of the 2.5-kb *Hin*dIII fragment within pL9.2 (see diagram Figure 4) was detected by the dideoxy sequencing method of Sanger *et al.* (1977). The fragment was subcloned into M13mp19 in both orientations, and sets of overlapping deletions were constructed using the methods of Henikoff (1984). The unique *Kpn*I and *Sal*I sites in the mp19 polylinker were cut, and the insert was progressively deleted from the *Sal*I end by exonuclease III. Fragments to the right and left of the *Hin*dIII sites in pL9.2 were sequenced after subcloning various overlapping restriction fragments into pTZ 18R or 19R vectors (Pharmacia).

Northern blots

Total RNA was isolated from mouse tissues by guanidine hydrochloride extraction (Cox, 1968; Strohman *et al.*, 1977), and from mouse L cells by guanidinium isothiocyanate extraction according to Maniatis *et al.* (1982). Poly(A)$^+$ RNA was purified by oligo(dT) chromatography. Some poly(A)$^+$ RNA from intestine, submaxillary gland, spleen, kidney and testis was a gift from Richard Meehan, MRC Western General Hospital, Edinburgh. RNA samples (5 μg) were fractionated through 1.5% agarose/2.2 M formaldehyde gels in 10 mM phosphate buffer (pH 6.5) and blotted onto nitrocellulose filters in 10 × SSC. After hybridization (16 h at 68°C) the filters were washed in 4 × SSC, 2 × SSC, 0.2 × SSC for 30 min at 65°C and autoradiographed at −70°C with intensification. Sizes of RNA components were estimated from mobilities relative to the major rRNA bands.

Hybridization probes

pL9.2, pL9.3 and pL9.5 are genomic *Eco*RI fragments subcloned into pUC9 as described (Bird *et al.*, 1985). The A and C probes were obtained by subcloning the 1.3- and 1.5-kb *Eco*RI–*Xho*I fragments of pL9.2. In the Northern blot experiments 100 ng of nick-translated plasmid were used per hybridization. Single-stranded probes were obtained from the M13 subclones of the *Hin*dIII fragment (see above). Second strand synthesis was primed using the 17-mer universal sequencing primer and the strands were uniformly labelled according to Meinkoth and Wahl (1984).

cDNA library screening

Clones (5×10^5) were screened from a C57 black 8.5-day mouse embryo cDNA library cloned into the *Eco*RI site of λ gt10, obtained from B.Hogan (MRC, London) and K.Fannher (Biogen). Seventeen positive clones were then rescreened individually. All the screening experiments were performed with labelled restriction fragments purified from low-melting point agarose gels (Feinberg and Vogelstein, 1984).

cDNA subcloning and sequencing

Lambda clones 10, 16, 17 and 6 were *Eco*RI digested and subcloned into M13mp19 or pTZ18R vectors. The recombinant plasmids containing the cDNA-5′ *Eco*RI fragments in both orientations were selected and sequenced by the dideoxy method as above.

S1 mapping

A 265-bp *Taq*I–*Sma*I or *Taq*I–*Xma*I fragment was end labelled at either the *Taq*I end or the *Xma*I end with [γ-^{32}P]ATP (3000 Ci/mmol, Amersham) and polynucleotide kinase. Strand separation was achieved on a 7% polyacrylamide gel, although resolution of the two strands was not optimal. By manipulating the order of restriction enzyme cutting and dephosphorylation, only one end of the double-stranded fragment was labelled for each set of experiments. Single-stranded probes were annealed to 50 μg of total liver RNA or 20 μg of yeast transfer RNA and then digested with S1 nuclease using the method of Favaloro *et al.* (1980). The annealing temperature was 62°C. S1-resistant products were analysed on polyacrylamide–urea gels and autoradiographed at −70°C with intensification.

Acknowledgements

We are grateful to Richard Meehan for the gift of mouse RNA, Brigid Hogan (National Institute for Medical Research, London) and K.Fannher (Biogen) for the mouse embryo cDNA library, Joe Lewis for help with part of the cDNA sequencing, and Anne Deane for typing the manuscript. P.L. was an EMBO Long Term Fellow on leave of absence from Centro di Genetica Evoluzionistica, CNR. The work was supported by the Medical Research Council, London.

References

Bird,A.P. (1986) *Nature*, **321**, 209–213.
Bird,A., Taggart,M., Fromer,M., Miller,O.J. and Macleod,D. (1985) *Cell*, **40**, 91–99.
Bird,A.P., Taggart,M.H., Nicholls,R.D. and Higgs,D.R. (1987) *EMBO J.*, **6**, 999–1004.
Brown,W.R.A. and Bird,A.P. (1986) *Nature*, **322**, 477–481.
Cooper,D.N., Taggart,M.H. and Bird,A.P. (1983) *Nucleic Acids Res.*, **11**, 647–658.
Cox,R. (1968) *Methods Enzymol.*, **65**, 120–129.
Crouse,G.F., Leys,E.L., McEwan,R.N., Frayne,E.G. and Kellems,R.E. (1985) *Mol. Cell Biol.*, **5**, 1847–1858.
Dynan,W. (1986) *Trends Genet.*, **6**, 196–197.
Farnham,P.J., Abrams,J.M. and Schimke,R.T. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 3978–3982.
Favaloro,J., Treissman,R. and Kamen,R. (1980) *Methods Enzymol.*, **65**, 718–749.
Feinberg,A. and Vogelstein,B. (1984) *Anal. Biochem.*, **137**, 266–267.
Giguere,V., Isobe,K.-I. and Grosveld,F. (1985) *EMBO J.*, **4**, 2017–2024.
Henikoff,S. (1984) *Gene*, **28**, 351–359.
Kadonaga,J., Jones,K. and Tjian,R. (1986) *Trends Biochem. Sci.*, **11**, 20–23.
Kolsto,A.-B., Kollias,G., Giguere,V., Isobe,K.-I., Prydz,H. and Grosveld,F. (1986) *Nucleic Acids Res.*, **14**, 9667–9678.
Kozak,M. (1986) *Cell*, **47**, 481–483.
McKeon,C., Ohkubo,H., Pastan,I. and de Crombrugghe,B. (1982) *Cell*, **29**, 203–210.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory Press, New York, pp. 196–198.
Meinkoth,J. and Wahl,G. (1984) *Anal. Biochem.*, **138**, 267–284.
Melton,D.W., Lorecki,D.S., Brennand,J. and Caskey,C.T. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 2147–2151.
Melton,D.W., McEwan,C., McKie,A.B. and Reid,A.M. (1986) *Cell*, **44**, 319–328.
Mitchell,P.J., Carothers,A.M., Han,J.H., Harding,J.D., Kas,E., Venolia,L. and Chasin,L. (1986) *Mol. Cell. Biol.*, **6**, 425–440.
Mount,S.M. (1982) *Nucleic Acids Res.*, **10**, 459–472.
Reynolds,G., Basu,S., Osborne,T., Chin,D., Gil,G., Brown,M., Goldstein,J. and Luskey,K. (1984) *Cell*, **38**, 275–285.
Saffer,J.D. and Singer,M.F. (1984) *Nucleic Acids Res.*, **12**, 4769–4788.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Sharp,P.A. (1981) *Cell*, **23**, 643–646.
Stein,R., Sciaky-Gallili,N., Razin,A. and Cedar,H. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 2423–2426.
Strohman,R., Moss,P., Micou-Eastwood,J., Spector,P., Prebyla,A. and Patterson,B. (1977) *Cell*, **10**, 265–273.
Toniolo,D., D'Urso,M., Martini,G., Persico,M., Tufano,V., Battistuzzi,G. and Luzzato,L. (1985) *EMBO J.*, **3**, 1987–1995.
Tykocinski,M.L. and Max,E.C. (1984) *Nucleic Acids Res.*, **12**, 4385–4396.
Williams,T.J. and Fried,H. (1986) *Mol. Cell. Biol.*, **6**, 4558–4569.

Received on April 22, 1987; revised on June 1, 1987