The EMBO Journal vol.7 no.7 pp.2035–2041, 1988

Four classes of mRNA are expressed from the mouse *int-2* gene, a member of the FGF gene family

Suzanne L.Mansour[1] and Gail R.Martin

Department of Anatomy, University of California at San Francisco, San Francisco, CA 94143, USA

[1]Present address: Department of Biology, University of Utah, Salt Lake City, UT 84112, USA

Communicated by P.Gruss

Mouse embryos at 7.5 days of gestation and endodermal cells derived from embryonal carcinoma cells each express four *int-2* mRNAs of similar size and relative abundance. To determine their structure and coding potential, we prepared a cDNA library from endoderm mRNA and isolated several *int-2* cDNAs. Structural analysis of these cDNAs combined with Northern blot hybridization and primer extension analyses of *int-2* mRNA revealed that the differences between the mRNAs are generated through the use of two alternate transcriptional start sites and two alternate polyadenylation sites. All four mRNAs share a common core sequence that encodes a protein with amino acid similarity to fibroblast growth factor.

Key words: int-2/fibroblast growth factor/teratocarcinoma cells/endodermal cells/early mouse embryos

Introduction

Interest in the *int-2* gene has been stimulated by the recent finding that it is one of a growing family of genes related to the fibroblast growth factors (FGFs). At present there are five members of this family known, each of which shares ~40–60% amino acid sequence similarity with the others. In addition to the basic and acidic FGFs (Abraham *et al.*, 1986a,b; Jaye *et al.*, 1986), and *int-2* (Dickson and Peters, 1987), the family now includes hst/KS3, a transforming gene initially identified by transfection of murine NIH 3T3 cells with DNA samples from a variety of human cancerous and non-cancerous tissue (Delli Bovi *et al.*, 1987; Taira *et al.*, 1987; Yoshida *et al.*, 1987) and FGF-3, a gene isolated from a human bladder carcinoma (Zhan *et al.*, 1988). *In vivo*, basic and acidic FGFs are known to have a variety of mitogenic functions, including angiogenic activity (reviewed by Gospodarowicz *et al.*, 1986), but nothing is known as yet about the normal functions of the three recently identified members of the FGF gene family. However, the fact that expression of the *int-2* gene has been detected in the mouse embryo during the early post-implantation period of development, but not at late stages of gestation or in the adult, suggests that it plays some role in normal embryogenesis (Jakobovits *et al.*, 1986; Wilkinson *et al.*, 1988).

The *int-2* gene was first identified as a mouse genomic sequence adjacent to integrated proviruses of the mouse mammary tumor virus (MMTV) in a large proportion of virus-induced tumors in certain mouse strains (Peters *et al.*, 1983; Dickson *et al.*, 1984). The finding that *int-2* RNA is present in mammary tumors containing proviral inserts in the *int-2* genomic region, but is undetectable in non-neoplastic mammary tissue, is consistent with the idea that activation of the *int-2* gene following proviral insertion can result in mammary carcinogenesis (Dickson *et al.*, 1984; Peters *et al.*, 1986). The fact that *int-2* is a member of a family of known growth factor genes suggests mechanisms by which unscheduled expression of *int-2* in mammary tissue could result in tumor formation.

The structure of the *int-2* gene and the sequence of its protein product have been deduced from studies of the *int-2* genomic region and of cDNA clones of *int-2* mRNA derived from a mammary carcinoma (Moore *et al.*, 1986). From these data it was concluded that the *int-2* gene extends over a region of <8 kb of genomic DNA, and contains three exons. However, these data had to be interpreted with some caution, because the cDNA that was sequenced was derived from a tumor in which the provirus had integrated such that transcription of the *int-2* gene in this tumor initiated within the proviral LTR rather than *int-2* genomic DNA. It is thus possible that the provirus may have created pseudo-exon sequences by subverting the normal transcription and/or RNA splicing patterns. Moreover, it was reported that the pattern of *int-2* transcripts detected depended on which tumor served as the source of *int-2* RNA and which particular probe was employed, suggesting that opportunities might exist for creating multiple *int-2* transcripts and that the structure of the gene might be more complex than that reported.

It became clear this was indeed the case when cells in which *int-2* is normally expressed were identified. Four species of *int-2* RNAs were detected in normal embryos at 7.5 days of gestation, and a similar pattern of expression was observed in teratocarcinoma stem cells that had differentiated to primitive endoderm (Jakobovits *et al.*, 1986). These observations raised the possibility that in non-neoplastic cells the mRNA expressed from the gene encodes a protein product(s) whose sequence is different from that of the product of the *int-2* tumor cDNA. In order to explore this question, and to define the structure of the gene in embryonic cell types, we isolated and sequenced *int-2* clones from a cDNA library prepared from mRNA isolated from a teratocarcinoma-derived endodermal cell population. The results of this analysis indicate that the four different transcripts are generated by the use of two different transcription start sites and two different polyadenylation sites. However, it appears that the differences among these four transcripts are all localized to their untranslated portions, and that all four transcripts code for the same product, an FGF-related protein of ~27 000 daltons.

Results

Characterization of int-2 cDNA clones derived from endoderm RNA

In a previous study of teratocarcinoma-derived endoderm RNA, a 598-bp probe derived from the 3′ portion of the *int-2* transcription unit was found to hybridize to four *int-2* transcripts (Jakobovits *et al.*, 1986), which we now estimate to be 2.9, 2.6, 1.7 and 1.4 kb in length (see below). To determine the structure of these transcripts and to compare them with the *int-2* tumor transcript described by Moore *et al.* (1986), we constructed several cDNA libraries in λgt10, starting with cytoplasmic poly(A)$^+$ RNA from PSA-1 endodermal cells. The 598-bp *int-2* probe was used to screen a total of 2.4 × 10^6 independent recombinant phage. After three rounds of plaque hybridization and purification we obtained nine phage that contained *int-2* sequences, ranging in size from ~800 to 2500 bp. The cDNA inserts were excised from the phage, subcloned into a plasmid vector and subjected to extensive restriction analysis (data not shown). The restriction maps of all the cDNAs could be aligned with the published composite tumor cDNA map, suggesting that *int-2* mRNA in PSA-1 endoderm is spliced in the same manner as the *int-2* tumor mRNA previously described. This conclusion was confirmed by DNA sequence analysis of the appropriate regions in several of the cDNAs (data not shown). There were, however, differences found between the endoderm and tumor cDNAs at their 5′ and/or 3′ ends (Figure 1).

With respect to their 3′ ends, the endoderm cDNAs could be grouped in two classes, one that appears to have the same 3′ end as the tumor cDNA (the 'long' class), and one that is ~1.1 kb shorter than the tumor cDNA (the 'short' class). Three of the endoderm cDNAs fell into the 'long' class, and six into the 'short' class. The 3′ end of one cDNA from each class was subjected to DNA sequence analysis. The results showed that the 3′ end of the 'long' cDNA was polyadenylated and mapped to position 7508 of the published *int-2* genomic sequence (Moore *et al.*, 1986), 21 bp downstream from the variant polyadenylation signal AATACA which is found in a very small percentage of eukaryotic messenger RNAs (Wickens and Stephenson, 1984). Its 3′ end is thus identical to that of the *int-2* tumor cDNA. The 3′ end of the 'short' cDNA was also found to be polyadenylated and mapped to position 6404 of the genomic *int-2* sequence, 20 bp downstream from an atypical polyadenylation signal, ATTAAA, found in ~12% of eukaryotic mRNAs (Wickens and Stephenson, 1984). This site is in the apparently untranslated portion of the last exon of the tumor cDNA, 1104 bp upstream from the polyadenylation site of the 'long' transcripts. The presence of two alternative polyadenylation sites 1.1 kb apart accounts for the difference between the 'long' (2.9 and 2.6 kb) and 'short' (1.7 and 1.4 kb) classes of *int-2* mRNA detected by Northern analysis of endoderm RNA (Jakobovits *et al.*, 1986, and below).

With respect to their 5′ ends, the nine endoderm cDNAs again could be grouped into two classes (see Figure 1), one

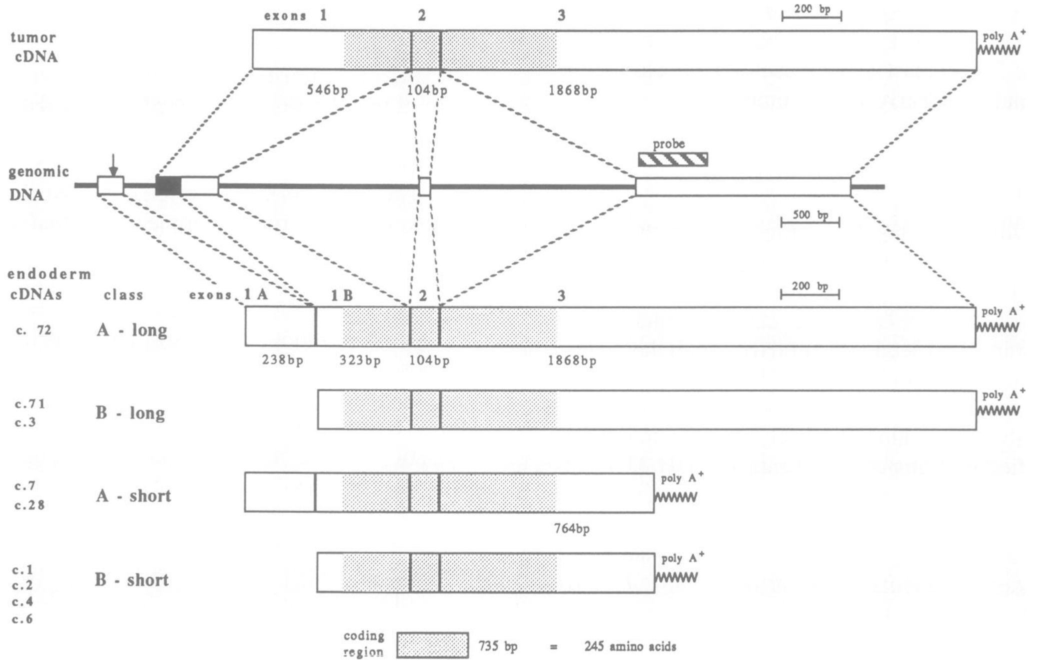

Fig. 1. Structure of *int-2* cDNA clones derived from endoderm RNA. The bold horizontal line in the upper portion of the diagram represents 7 kb of *int-2* genomic DNA, and the regions that are transcribed into mRNA in endodermal cells are delineated by open boxes. The cross-hatched box represents the 598-bp *Bam*HI/*Eco*RI genomic subclone that was used as a probe in the isolation of the cDNA clones. The bottom part of the figure illustrates the four classes of *int-2* cDNAs that were identified in this study. The designations of the individual clones that fell into each of the classes are listed on the left (e.g. clone c.72 was the only example of class A-long that was isolated). Exon 1A in clone c.28 spanned the region from positions 938 to 1175 in the genomic sequence published by Moore *et al.* (1986); the other class A clones did not extend quite as far 5′. In all class A clones exon 1A was joined to exon 1B at position 1671. The 5′ boundary of the class B clones is shown at position 1678, although none of the class B clones actually extended this far, and some of the class B clones lacked any exon 1B sequences; the conclusion that the mRNAs represented by the class B clones contain exon 1B and extend to this position is based on Northern blot hybridization experiments and primer extension analyses (see text). In all clones that contained it, exon 1B extended 3′ to genomic position 1993, and was joined to exon 2 at position 3737. In all clones, exon 2 extended to position 3840 and was joined to exon 3 at position 5641. The 'long' clone that was analyzed terminated at genomic position 7508, whereas the 'short' clone terminated at position 6404. The top part of the figure illustrates the *int-2* tumor cDNA described by Moore *et al.* (1986). The vertical arrow above the genomic DNA indicates the position at which the MMTV-provirus was inserted in the genomic DNA of the tumor cells from which the cDNA clone was derived. The region of genomic DNA that was not transcribed in the endodermal cells, but was found to be part of the transcription unit in the tumor cells, is delineated by a black box.

that began in the genomic region upstream of the first exon of the tumor cDNA (class A), and one that began within the first or second exon defined by the tumor cDNA (class B). Of the three cDNA clones that fell into class A, one had a 'long' 3′ end (clone c. 72) and two had a 'short' one (clones c.7 and 28). Sequence analysis of the 5′ end of each of these three clones revealed a new upstream exon, not found in the tumor cDNA. In the case of clone c.28, which extends the farthest 5′, this exon spans the region between genomic positions 938 and 1175. This exon is joined to sequences in the first exon of the tumor cDNA at genomic position 1671. The same splice junction was found in clones c.72 and c.7, which begin at positions 941 and 961 respectively. To avoid confusion, we will continue to refer to the first exon of the tumor cDNA as exon 1, the new first exon will be designated exon 1A and the second exon found in these endoderm cDNA clones, which is co-linear with part of the tumor cDNA exon 1, will be referred to as exon 1B (see Figure 1). The remaining six cDNA clones were arbitrarily placed in class B, because they did not contain any exon 1A sequences; however, any of them might be clones of a class A mRNA that was not fully reverse transcribed during the construction of the cDNA library. Evidence that a population of class B mRNAs does exist is presented below.

Structure of int-2 mRNAs in teratocarcinoma cells

The cDNA analysis suggested that all of the *int-2* mRNAs in PSA-1 endodermal cells share a common core region comprising exons 1B, 2 and part of exon 3. The differences between the various cDNAs appeared to be confined to their 5′ and 3′ ends by virtue of the optional inclusion of exon 1A at the 5′ end and/or the rest of exon 3 at the 3′ end. In order to establish exactly which combination of sequences was found in each particular mRNA, we carried out the following Northern blot hybridization experiment. Five samples, each containing 7 μg of poly $(A)^+$ RNA from PSA-1 embryoid bodies, were fractionated by gel electrophoresis, transferred to nylon membranes and hybridized with the *int-2* probes illustrated in Figure 2. Also included in this analysis was a sample containing a set of marker RNAs that could be detected by hybridization with a phage λ DNA probe (Figure 2, lane M).

As expected, probe D, the 598-bp genomic fragment that is derived from the 5′ portion of exon 3, and which was used for screening the cDNA libraries, hybridized to four transcripts (Figure 2, lane D). By comparison with the RNA standards, we now estimate these transcripts to be ~2.9, ~2.6, ~1.7 and ~1.4 kb in length. Similar results were obtained with probes B, which contains 303 bp of the 323-bp exon 1B as well as downstream intron sequences, and C, which contains the 104-bp exon 2 and flanking intron sequences (Figure 2, lanes B, C). The relative intensities of the signals generated with these three different probes were roughly proportional to the amount of exon sequences they contained. Thus, probe C gave the weakest signal.

In contrast to the results obtained with probes B, C and D, probe E hybridized to only the two larger of the four *int-2* transcripts (Figure 2, lane E). This probe consists of sequences at the 3′ end of the third exon, downstream of the first of the two potential polyadenylation signals found in the last exon of the *int-2* gene. The finding that this probe hybridizes to the pair of 'long' transcripts (2.9 and 2.6 kb), but not the pair of 'short' ones (1.7 and 1.4 kb) confirms the conclusion drawn from our analysis of *int-2* cDNA clones that both polyadenylation signals are used when *int-2* is expressed in teratocarcinoma-derived endodermal cells. Moreover, the observation that probes B, C and D detect the pairs of 'long' and 'short' transcripts in roughly equal abundance suggests that these alternative signals are used with roughly equal frequency.

A second probe, A, was also found to hybridize to only two of the four *int-2* transcripts. However, in contrast to the results obtained with probe E, probe A hybridized to the longer transcript of each pair (2.9- and 1.7-kb transcripts), and did not hybridize to the shorter transcript of either pair (2.6- or 1.4-kb transcripts; Figure 2, lane A). This conclusion was confirmed by rehybridizing the sample in lane A with probe B, to demonstrate that it was indeed the 2.9- and 1.7-kb transcripts that were detected by probe A (data not shown). Probe A contains the 238-bp exon 1A as well as 391 bp of upstream sequences and 46 bp of downstream sequences in the intron between exons 1A and 1B. These data indicate that exon 1A is present only in the 2.9-

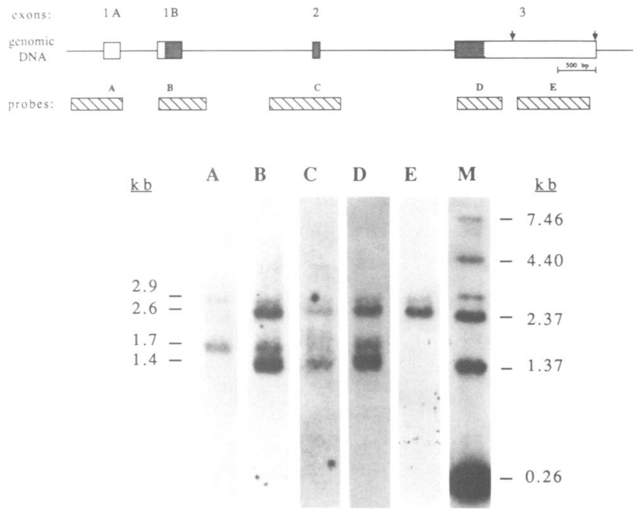

Fig. 2. Structure of *int-2* mRNAs as determined by Northern blot hybridization analysis. The top part of the figure illustrates the *int-2* genomic DNA; boxes delineate the positions of the four exons found in the class A-long clones. The stippled areas of the boxes indicate the extent of the presumed protein-coding sequence. The vertical arrows above exon 3 indicate the positions of the polyadenylation signals used in the 'short' and 'long' classes of clones. Below the genomic DNA, cross-hatched boxes represent the DNA probes used in this analysis. Probe A: a 675-bp *Xmn*I/*Stu*I genomic fragment that contains the 238-bp exon 1A as well as 391 bp of upstream sequences and 46 bp of downstream sequences; probe B; a 647-bp *Bss*HII/*Sst*I genomic fragment that contains 303 bp of the 323-bp exon 1B and downstream sequences; probe C: a 941-bp *Hin*cII/*Pst*I genomic fragment that contains the 104 bp exon 2; probe D: a 598-bp *Bam*HI/*Eco*RI genomic fragment consisting of exon 3 sequences; probe E: a 952-bp *Pvu*II/*Hin*dIII fragment derived from one of the 'long' cDNA clones. The bottom part of the figure shows the results of a Northern blot hybridization analysis of poly$(A)^+$ RNA from PSA-1 embryoid bodies using different *int-2* probes. Five samples, each containing 7 μg of poly $(A)^+$ RNA, were fractionated by gel electrophoresis and transferred to a nylon membrane. The lanes containing the individual samples were separated, and each was hybridized with one of the probes shown above. The particular probe used is designated above each lane. An additional sample included in the experiment was a RNA ladder containing a set of mol. wt standards, each of which contains a 154-base sequence of phage λ at its 3′ end. This lane was hybridized with a nick-translated probe for phage λ (**lane M**). The sizes of the marker RNAs are shown on the right, and the estimated sizes of the *int-2* transcripts are shown on the left.

and 1.7-kb transcripts, and that these two species represent a very small fraction of the *int-2* RNA present in endoderm. In view of this observation, it is somewhat surprising that three out of the nine clones we obtained were cDNA copies of these class A *int-2* mRNAs.

Since the *int-2* transcript represented by the tumor-derived cDNA clones previously analyzed (Moore *et al.*, 1986) contained ~220 bp of the genomic region that lies between exons 1A and 1B of the *int-2* endodermal cDNAs studied here, we sought to determine whether sequences from that region might be contained in any of the teratocarcinoma-derived mRNAs. An intron 1 probe (the region between probes A and B in Figure 2) containing only 20 bp of exon 1B sequences was hybridized to a sample of the PSA-1 embryoid body RNA. The only transcript to which this probe clearly hybridized was much larger than any of the *int-2* transcripts detected by probes A−E (data not shown), and was presumably an unspliced mRNA precursor. These data suggest that little, if any, of this region is included in the mRNAs expressed in cultures of differentiating PSA-1 teratocarcinoma cells.

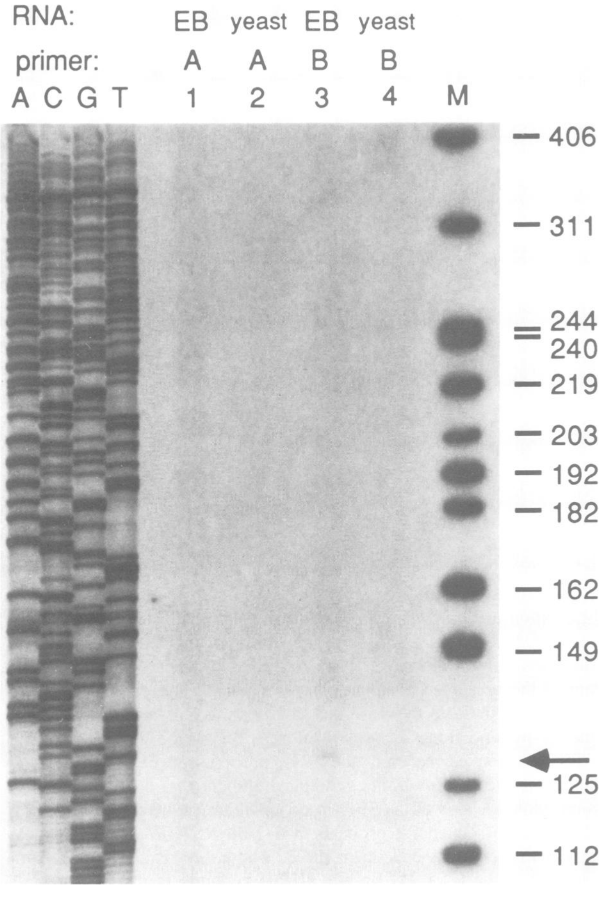

Fig. 3. 5′-End mapping of *int-2* mRNA by primer extension. 5′ end-labeled primers (**A**: specific for exon 1A, **lanes 1** and **2**; **B**: specific for exon 1B, **lanes 3** and **4**) were annealed to 10 μg of embryoid body (EB) mRNA (**lanes 1** and **3**) or yeast RNA (**lanes 2** and **4**). The primers were extended with reverse transcriptase, fractionated on a 6% polyacrylamide/8 M urea sequencing gel, and subjected to autoradiography for 10 days with an intensifying screen. Markers prepared by sequencing *int-2* genomic DNA with primer A are shown in the lanes marked **A**, **C**, **G**, **T**. End-labeled markers prepared from pBR322 DNA digested with *Msp*I are shown in **lane M** and their sizes in base pairs are indicated on the right. An arrow marks the position of the extension product used to determine the start site of the class B mRNAs.

Primer extension analysis

A comparison of the sizes of the class A cDNAs with the mRNAs that they represent indicates that those cDNA clones must be nearly full length. Thus, the 5′ ends of the mRNAs that contain exon 1A should map to the region just upstream of genomic position 938. For the class B *int-2* mRNAs, which apparently lack sequences from exon 1A, the results of the cDNA and Northern analyses suggested that they might actually be initiated downstream from the start of the class A mRNAs, near the border of the exon 1A/1B boundary. Neither of these presumed regions of transcription initiation contains the TATA sequence element that is a hallmark of many eukaryotic promoters.

In an attempt to determine precisely the transcription initiation sites, two 24-base synthetic oligonucleotide primers were prepared. Primer A is specific for exon 1A and primer B is specific for exon 1B. The primers were 5′ end-labeled, annealed to embryoid body mRNA or to yeast RNA and extended with reverse transcriptase. Figure 3 shows an autoradiogram of the gel-fractionated products from this reaction. Neither primer was extended when yeast RNA was used as the template (lanes 2 and 4). However, with embryoid body mRNA as a template, primer B was extended to a single site which, by comparison with the markers, is located 7 or 8 bp downstream from the splice junction between exons 1A and 1B at genomic position 1677 or 1678 (lane 3). This presumably represents the initiation site for the class B mRNAs. We expected that primer B would also hybridize with the class A mRNAs and be extended at least another 246 nucleotides beyond the class B start site. However, Northern blot analysis suggests that the class A mRNAs are ~10 times less abundant than the class B mRNAs (see Figure 2) and a signal 10 times less intense than the signal observed from the class B extension products could not have been detected. We were unable to detect the start of the class A mRNAs using primer A (lane 1), presumably for the same reason.

Discussion

Taken together, our data provide a description of the structure of the four *int-2* mRNAs found in endodermal cells derived from PSA-1 teratocarcinoma cells. The two most abundant mRNA species [class B, 2.6 kb (long) and 1.4 kb (short)] each consist of three exons (1B, 2 and 3), and the difference in their lengths is due to the use of alternative polyadenylation signals that are 1.1 kb apart in the genome. Data from primer extension studies suggest that transcription of these two RNA species is initiated at genomic position 1677 or 1678, presumably as a consequence of the activity of a promoter termed P_B. The two less abundant mRNA species [class A, 2.9 kb (long) and 1.7 kb (short)] are similar to the more abundant species, but each apparently contains an additional exon, 1A, at its 5′ end. Transcription of these mRNAs is presumably under the control of an upstream promoter, P_A, but the precise site of transcription initiation of these low-abundance mRNAs could not be determined. Our data concerning the 5′ end of the *int-2* gene are summarized in Figure 4.

It is likely that all four species of *int-2* mRNA are expressed in the same cell type. The PSA-1 endodermal cell population from which we isolated the four classes of cDNA clones consisted primarily of parietal endoderm-like cells (Grabel and Casanova, 1986). Furthermore, when Northern blot hybridization analyses were carried out on RNA from a presumably homogeneous population of parietal endoderm-like cells obtained by treating F9 embryonal carcinoma cells with retinoic acid and dibutyryl-cAMP (Strickland *et al.*, 1980), the results were identical to those obtained with RNA from PSA-1-derived endodermal cells (as in Figure 2, lanes A, B; data not shown). This indicates that the same four species of *int-2* RNAs, at the same relative abundance, are present in both PSA-1- and F9-derived endodermal cells.

```
 501  gaagcacagg atgtggggtg ggaccttttc tgatacctca ggaagtcctt
 551  cttctttgag gcccttggca gataaggaac tgaggtcaag agctcagggt
 601  ccagggagct attccctgaa cctggactcc ccctcagcct gccaggtggg
 651  catgcgggcg ccattgtggg atcttctttc ccttcggttc tgttgccctc  PA
 701  tgtatctcca cattcctccc cctctccact gctctcctcc ttctttctct
 751  tcagccaaat ctgtcatctg cctcacccac ctttatttct gtctctgcct
 801  agatttcctg ggctctaagg ctctgtgact ctattgtctc tgctcctatc
 851  tgtgcagtcc cctctctgga ccaccgctgg gacagaccac tcccacttaa

 901  ggtgtgcaaa cttcctgtgc taccttctgg tcagttaGGT AGCTTCGAGC

 951  TTAGAGTCCC TGGCCCGTTG GGCCACAAGA CTGGACGCCT GAGGACCGCT

1001  CTGCGGTGTG AGAAGCGACC TGCACGAACC CAACCCCGGG CTCCCCACGC
                                                               1A
1051  ACCGCCCTCC GCAATTGCTC TGGATTAGGC GCCGCCTTTC ATGAGCGTTT

1101  GTCAGCTAGA CTTCCCCGCA AGTTGTTTGC GCAGACATCA GTCATCCACA

1151  GTCCCATTGT GCACCCAGGA TTGATgtaag gcgggagggg gtgacagggc

1201  ctgggggcgg ttgcctttag gcctcgctct ataattttcc gaggcataat
1251  tggtctgggg gcgggggcgg ggcggggacc tttcagaggc gggaggggc
1301  tcagggcgca cggcggagga gcggcggccc gacggctctg gcccgggagc
1351  tgtgcgcagg cgacgcccgg cctgagtccc gcgcccccgc cagggaccac
1401  ggccgccttt tgttgtcaag cgccctttct tcagaactgt gttcggcaaa  PB
1451  gaaacacgac ccccattcct gggtgaaaat tcaaagtctt ctttctccat
1501  ctccctctcc tctttcttgt cttctctcca ctatctcttc ccctctcctg
1551  ctccacccgt ccttatctcc atctccacct ctctccacct ctccctgtct
1601  cccctcctct tcctccctcc cctgtctccc acctcttccc ggtcccccctc
                                  B
                                  ⇒
1651  tccctccctt cctctcctag CCACAGTCGA GCCGGCCTGG CGCGCGGGCG

1701  TGTGCTCCCA GCGCCGCGCC TTCGTGAGAC CCGCGCTGGC GCAGCAGCCG

1751  CTGCGGGCGG GCGCGATGCC GGGATGGGCC TGATCTGGCT TCTGCTGCTC

1801  AGCTTGCTGG AACCCAGCTG GCCAACTACG GGGCCCGGGA CGCGACTACG  1B

1851  ACGCGATGCG GGCGGCCGTG GTGGCGTTTA CGAGCACCTC GGCGGGGCGC

1901  CACGGCGCCG CAAGCTCTAC TGCGCTACCA AGTACCACCT CCAGCTGCAC

1951  CCAAGCGGCC GCGTGAACGG CAGCCTTGAG AACAGCGCCT ATAgtgagtg
```

Fig. 4. DNA sequence surrounding *int-2* exons 1A and 1B. The sequence of *int-2* genomic DNA between nucleotides 501 and 2000 is taken from Moore *et al.* (1986). Exon sequences as defined by analysis of cDNA clones are depicted in capital letters and designated 1A and 1B. The presumed promoter regions are designated P_A and P_B. The start site for the class B mRNAs is indicated with an arrow and the letter B. The potential binding sites for CTF/NF-1 (5′AGCCAA3′, Jones *et al.*, 1987) and Sp1 ($5'\frac{G}{T}\frac{G}{A}GG\frac{C}{A}G\frac{G}{T}\frac{G}{A}\frac{G}{A}\frac{C}{T}3'$, (Kadonaga *et al.*, 1986; J. Kadonaga, personal communication) are underlined. (The bases shown in the upper position in the sequence above are the preferred ones). Note that the potential SP1 binding sites around position 1600 are encoded by the strand complementary to the one shown in the figure. The major ORF beginning in exon 1B is underlined with a bold line, and the upstream, in-frame CTG is boxed. The 27 amino acid ORF in exon 1A is highlighted with a stippled underline.

The data described here are also consistent with some of the conclusions of a recent study of the *int-2* gene by Smith *et al.* (1988). In that study, the results of RNase protection assays provided evidence of the utilization of the same two polyadenylation sites that we identified by cDNA cloning and sequencing. Moreover, both studies are in agreement that the 3′ boundary of 'exon 1B', the 5′ and 3′ boundaries of 'exon 2' and the 5′ boundary of 'exon 3' are the same in all *int-2* transcripts.

However, with respect to the structures of the 5′ ends of the *int-2* mRNAs and the locations of the promoters that control expression of the *int-2* gene, the conclusions of the two studies differ markedly. From their data, Smith *et al.* concluded that F9-derived endodermal cells contain only two *int-2* transcripts, both with 'exon 1A' at their 5′ ends; thus they inferred that all transcription of the *int-2* gene in F9-derived endoderm occurs from a promoter domain they termed P1, which is the same as the one we have termed P_A. Moreover, on the basis of data from RNase protection assays and sequence analysis they concluded that alternative splice acceptor sites for the joining of exon 1A to exon 1B in these mRNAs exist at genomic positions 1671 and 1677. However, the results of our Northern blot hybridization experiments clearly demonstrate that these two major *int-2* transcripts lack exon 1A and intron 1 sequences in both PSA-1- and F9-derived endodermal cells (Figure 2, lane A and data not shown), and suggest that most transcription of the *int-2* gene in both cell types is controlled by a promoter domain that we have termed P_B. Our data from primer extension and cDNA cloning studies further suggest that genomic position 1677 or 1678 is the start site for transcription from P_B, rather than a splice acceptor site. The finding that both PSA-1- and F9-derived endodermal cells do contain small quantities of mRNAs with exon 1A (Figure 2, lane A and data not shown) further suggests that promoter $P1/P_A$ is active at a very low level in these cells.

One intriguing finding of Smith *et al.* is that a different pattern of *int-2* expression is observed in retinoic-acid-treated PCC4 embryonal carcinoma cells (RA-PCC4). Northern blot hybridization analysis demonstrated that these cells express four transcripts, presumably similar in size to those described here. However, in contrast to the situation in PSA-1- or F9-derived endodermal cells, it is the larger transcripts of the long and short classes that are most abundant in RA-PCC4. From the results of RNase protection assays, Smith *et al.* have drawn several conclusions about the structures of the 5′ ends of those RA-PCC4 transcripts, but in the absence of any corroborating data from Northern blot hybridization analysis it is difficult to evaluate the validity of their hypotheses. However, the important point is that the first exon in at least some of the RA-PCC4 mRNAs appears to be similar to the first exon of the *int-2* tumor cDNA previously described by Moore *et al.* (1986). These mRNAs therefore contain sequences located in the genomic region between the 1A and 1B exons that are not found in the major transcripts in PSA-1- or F9-derived endodermal cells. Thus transcription initiation in RA-PCC4 cells is controlled by the activity of a promoter, termed P2 by Smith *et al.*, that is different from the $P1/P_A$ and P_B promoters discussed above.

Taken together, the data from the two studies indicate that *int-2* gene expression can be controlled by at least three different promoters (see Figure 4): the $P1/P_A$ promoter, identified in both studies, which may control expression at a low level in the three cell types studied; the P_B promoter which we have identified, and which appears to be the most active promoter in PSA-1- and F9-derived endodermal cells and may also function in RA-PCC4 cells; and the P2 promoter identified by Smith *et al.* (1988), which is the most active one in RA-PCC4 cells and some *int-2* tumors but which may not function in PSA-1- or F9-derived endodermal cells. This promoter domain lies at the 5′ end of the region designated P_B in Figure 4. As yet, details of the structures of the *int-2* transcripts in embryonic cells are not available, and it thus remains to be determined which of these three promoters are active in embryos. The intriguing possibility exists that different combinations of *int-2* mRNA classes may be expressed in different cell types in the embryo. A hint that this may be the case comes from a comparison of our previous findings (Jakobovits *et al.*, 1986) with the data of Wilkinson *et al.* (1988). We showed that four transcripts similar in size and relative abundance to those detected in PSA-1- endodermal cells are expressed in mouse embryos at 7.5 days of gestation, suggesting that at the early post-implantation stages of development the same promoters that control *int-2* expression in PSA-1- and F9-derived endoderm are active. However, from the study by Wilkinson *et al.* it appears that in embryos at 9.5 days of gestation the pattern of expression of *int-2* is different, at least with respect to the relative abundance of the various size transcripts. Ultimately this issue will only be resolved by *in situ* hybridization studies of sections of mouse embryos employing probes that distinguish the different classes of *int-2* mRNAs.

At present the nature of the promoter elements that regulate the transcription of the class A and B *int-2* RNAs is unclear. Our primer extension analysis revealed a likely start site for the class B mRNAs. Size comparisons of class A cDNA clones with mRNAs detected by Northern blot analysis suggest that the clones are nearly full length, thus identifying the genomic region in which the actual 5′ end of the class A mRNAs is likely to be found. In neither case is a TATA box found in the sequences upstream from these regions of transcription initiation. There is, however, a half-site (5′AGCCAA3′) for the CAAT-binding transcription factor, CTF/NF-1 (Jones *et al.*, 1987) located upstream from the class A start site (see Figure 4). In addition, several binding sites for the transcription factor Sp1 (Kadonaga *et al.*, 1986) can be found in the region between exons 1A and 1B (see Figure 4), but it is not yet known whether any of these sites are involved in the regulation of *int-2* transcription. It is also uncertain what regulates transcript processing and why approximately half of the transcripts in each class extend through the upstream polyadenylation signal to be processed at the downstream site.

Analysis of the sequences that generate the differences among the four *int-2* mRNAs failed to reveal any obvious functional significance of these different transcripts. The 1.1 kb of additional sequence at the 3′ ends of the 'long' transcripts (2.9 and 2.6 kb) is 3′ to the translation termination signal, thus precluding the addition of any coding potential. Moreover, although inclusion of exon 1A in the class A transcripts provides 89 N-terminal amino acids of coding potential in frame with the major open reading frame (ORF) found in exons 1B, 2 and 3, the lack of an appropriate AUG translation initiation codon in this stretch of sequence would appear to preclude inclusion of this amino acid sequence in the *int-2* protein product. However, it has recently been reported that translation initiation on c-*myc* mRNA occurs naturally at both an AUG codon in exon 2 and at a CUG codon located upstream in exon 1 (Hann *et al.*, 1988). There is a CUG codon located upstream from and in frame with the AUG codon that begins the *int-2* major ORF (see Figure 4). If this codon were used to initiate translation of *int-2*, an additional 29 amino acids would be added to the 245 amino acids already encoded by the major FGF-related ORF. Nevertheless, all four *int-2* mRNAs would still encode the same protein product(s) because the CUG codon is located downstream from the exon 1A/1B border and thus is contained in all four *int-2* mRNAs. It should, however, be noted that exon 1A does contain a short ORF that could potentially encode a polypeptide of 27 amino acids. If this ORF were translated, it would provide the class A mRNAs with a unique function. Alternatively, the presence of this ORF on the class A mRNAs could contribute to the regulation of translation of the major ORF located downstream.

The most intriguing unresolved questions, however, concern the normal function of the *int-2* protein product and the role it plays in the development of the early embryo. Its similarity to FGF suggests that it may function as a growth regulatory molecule. However, it is also possible that it functions as an inducer of differentiation, in view of recent studies suggesting that in the *Xenopus* embryo basic FGF can act as a morphogen, inducing differentiation of dorsal ectoderm to mesoderm (Slack *et al.*, 1987; Kimmelman and Kirschner, 1987). However, there is no information at present on the extent to which *int-2* activity is similar to that of FGF. The recent work of Wilkinson *et al.* (1988) has defined cell types in the mouse embryo in which *int-2* RNA is expressed, thus confirming that *int-2* expression is highly restricted both temporally and spatially during embryogenesis. However, the significance of *int-2* expression in such diverse cell populations as extra-embryonic parietal endoderm, mesodermal cells migrating through the primitive streak, neuroepithelium adjacent to developing otocysts and the endoderm of the pharyngeal pouches remains obscure. Clearly the possibility exists that *int-2* may have several different functions in embryogenesis, and an understanding of these functions may lead to new insights into the mechanisms that govern morphogenesis.

Materials and methods

Cell cultures

PSA-1 teratocarcinoma stem cells were cultured in aggregates and induced to form embryoid bodies as described by Martin *et al.* (1977). To obtain an endodermal cell population, the embryoid bodies were allowed to attach to collagen-coated tissue culture dishes, and the endodermal cells that migrated along the substratum and away from the embryoid bodies were separated from the cells that remained in the embryoid bodies by treatment with dispase and collagenase, as described by Grabel and Martin (1983).

RNA isolation and Northern blot hybridization analysis

Cytoplasmic RNA was obtained from embryoid bodies or isolated endodermal cell populations using the method of Berk and Sharp (1977). Total cellular RNA was prepared using the LiCl precipitation method of Cathala *et al.* (1983). Polyadenylated RNA was purified from these preparations by the method of Aviv and Leder (1972).

Poly(A)$^+$ RNAs were fractionated by electrophoresis on denaturing 1% agarose/2.2 M formaldehyde gels at room temperature. Following transfer of the RNA samples to GeneScreen membranes the blots were hybridized with DNA probes labeled by nick translation with [^{32}P]-dNTPs as previously described by Joyner *et al.* (1985). mRNA sizes were estimated by comparison with RNA size standard markers purchased from Bethesda Research Laboratories (BRL, Bethesda, MD).

cDNA library construction

Two cDNA libraries were constructed using cytoplasmic RNA from an isolated endodermal cell population derived from PSA-1 embryoid bodies. Starting with 5 μg of poly (A)$^+$ RNA, the first library was constructed by following the protocol of Huynh *et al.* (1985), with the modification of Gubler and Hoffman (1983) for the second strand reaction. The second library was made starting with 10 μg of poly (A)$^+$ RNA, and included several steps designed to maximize the production of full-length cDNA (N.Crawford, personal communication). The cDNAs (>500 bp, library 1; >2000bp, library 2) were ligated to λgt10 arms [prepared according to Huynh *et al.* (1985) or purchased from Promega, Madison, WI] and packaged using Gigapack extract from Stratagene (San Diego, CA). In both cases the recombinant phage were screened prior to amplification according to the method of Maniatis *et al.* (1982) using DNA probes nick-translated to a specific activity of $>2 \times 10^8$ c.p.m./μg. The two libraries yielded nine independent *int-2* cDNA clones from a total of 2.4×10^6 recombinant phage screened.

Analysis of cDNAs

cDNA inserts in phage were subcloned into Bluescript vectors (Stratagene) and subjected to restriction analysis (Maniatis *et al.*, 1982). Selected portions of the cDNAs were sequenced by subcloning the appropriate fragments into Bluescript vectors and either rescuing single strands for analysis by the dideoxynucleotide sequencing method of Sanger *et al.* (1977) or denaturing double-stranded miniprep DNA (C.Thacker, personal communication) and following the sequencing protocol of Tabor and Richardson (1987) that employs a modified T7 DNA polymerase (US Biochemicals, Cleveland, OH).

Primer extension analysis

Two synthetic 24-base oligonucleotides were synthesized using an Applied Biosystems machine.

The sequence of primer A is 5'TGGATGACTGATGTCTGCGCAAAC3'.
The sequence of primer B is 5'AAGCTGAGCAGCAGAAGCCAGATC3'.
The oligonucleotides were purified following preparative electrophoresis on a 20% polyacrylamide/8 M urea gel and 5' end-labeled using [γ-^{32}P] ATP and T4 polynucleotide kinase. The labeled oligonucleotides were annealed to 10 μg of embryoid body poly(A)$^+$ RNA at 51°C for 6 h and then extended using 10 U of avian myeloblastosis virus reverse transcriptase (Seikigaku, St Petersburg, FL) as described by McKnight and Kingsbury (1982). The products of the reaction were fractionated by electrophoresis through a 6% acrylamide/8 M urea sequencing gel.

Acknowledgements

We thank Clive Dickson and Gordon Peters for kindly providing a computer diskette with the complete *int-2* genomic sequence. We are also grateful to John Scott for excellent technical assistance in carrying out the Northern blot hybridization analysis, Martin Carroll for his help in characterizing several of the cDNA clones, and Mario Capecchi for providing S.L.M. with laboratory space in which to complete this work. This work was supported by grant CD-319 from the American Cancer Society. S.L.M. was initially supported by NIH-postdoctoral training grant 2T32 CA09043 and subsequently by a postdoctoral fellowship from the American Cancer Society.

References

Abraham,J.A., Mergia,A., Whang,J.L., Tumulo,A., Friedman,J., Hjerrild,K.A., Gospodarowicz,D. and Fiddes,J.C. (1986a) *Science*, **233**, 545–548.

Abraham,J.A., Whang,J.L., Tumulo,A., Mergia,A., Friedman,J., Gospodarowicz,D. and Fiddes,J.C. (1986b) *EMBO J.*, **5**, 2523–2528.

Aviv,H. and Leder,P. (1972) *Proc. Natl. Acad. Sci. USA*, **69**, 1408–1412.

Berk,A.J. and Sharp,P.A. (1977) *Cell*, **12**, 721–732.

Cathala,G., Savouret,J.-F., Mendez,B., West,B.L., Karin,M., Martial,J.A. and Baxter,J.D. (1983) *DNA*, **2**, 329–335.

Delli Bovi,P., Curatola,A.M., Kern,F.G., Greco,A., Ittmann,M. and Basilico,C. (1987) *Cell*, **50**, 729–737.

Dickson,C. and Peters,G. (1987) *Nature*, **326**, 833.

Dickson,C., Smith,R., Brookes,S. and Peters,G. (1984) *Cell*, **37**, 529–536.

Gospodarowicz,D., Neufeld,G. and Schweigerer,L. (1986) *Mol. Cell. Endocrinol.*, **46**, 187–204.

Grabel,L.B. and Martin, G.R. (1983) *Dev. Biol.*, **95**, 115–125.

Grabel,L.B. and Casanova,J.E. (1986) *Differentiation*, **32**, 67–73.

Gubler,U. and Hoffman,B.J. (1983) *Gene*, **25**, 263–269.

Hann,S.R., King,M.W., Bentley,D.L., Anderson,C.W. and Eisenman,R.N. (1988) *Cell*, **52**, 185–195.

Huynh,T.V., Young,R.A. and Davis,R.W. (1985) In Glover,D.M. (ed.), *DNA Cloning—A Practical Approach. Vol. I.* IRL Press, Oxford, pp. 49–78.

Jakobovits,A., Shackleford,G., Varmus,H.E. and Martin,G.R. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 7806–7810.

Jaye,M., Howk,R., Burgess,W., Ricca,G.A., Chiu,I.-M., Ravera,M.W., O'Brien,S.J., Modi,W.S., Maciag,T. and Drohan,W.N. (1986) *Science*, **233**, 541–545.

Jones,K.A., Kadonaga,J.T., Rosenfeld,P.J., Kelly,T.J. and Tjian,R.T. (1987) *Cell*, **48**, 79–89.

Joyner,A.L., Kornberg,T., Coleman,K., Cox,D. and Martin,G.R. (1985) *Cell*, **43**, 29–37.

Kadonaga,J.T., Jones,K.A. and Tjian,R. (1986) *Trends Biochem.*, **11**, 20–23.

Kimmelman,D. and Kirschner,M. (1987) *Cell*, **51**, 869–877.

Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Martin,G.R., Wiley,L.M. and Damjanov,I. (1977) *Dev. Biol.*, **61**, 230–244.

McKnight,S.L. and Kingsbury,R. (1982) *Science*, **217**, 316–324.

Moore,R., Casey,G., Brookes,S., Dixon,M., Peters,G. and Dickson,C. (1986) *EMBO J.*, **5**, 919–924.

Peters,G., Brookes,S., Smith,R. and Dickson,C. (1983) *Cell*, **33**, 369–377.

Peters,G., Lee,A.E. and Dickson,C. (1986) *Nature*, **320**, 628–631.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5436–5467.

Slack,J.M.W., Darlington,B.G., Heath,J.K. and Godsave,S.F. (1987) *Nature*, **326**, 197–200.

Smith,R., Peters,G. and Dickson,C. (1988) *EMBO J.*, **7**, 1013–1022.

Strickland,S., Smith,K.K. and Marotti,K.R. (1980) *Cell*, **21**, 347–355.

Tabor,S. and Richardson,C.C. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4767–4771.

Taira,M., Yoshida,T., Miyagawa,K., Sakamoto,H., Terada,M. and Sugimura,T. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 2980–2984.

Wickens,M. and Stephenson,P. (1984) *Science*, **226**, 1045–1051.

Wilkinson,D.G., Peters,G., Dickson,C. and McMahon,A.P. (1988) *EMBO J.*, **7**, 691–695.

Yoshida,T., Miyagawa,K., Odagiri,H., Sakamoto,H., Little,P.F.R., Terada,M. and Sugimura,T. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 7305–7309.

Zhan,X., Hu,X. and Goldfarb,M. (1988) *Mol. Cell. Biol.*, in press.

Received on March 24, 1988; revised on April 18, 1988