The EMBO Journal vol.4 no.11 pp.2905–2909, 1985

The nucleotide sequence of the human *int*-1 mammary oncogene; evolutionary conservation of coding and non-coding sequences

Albert van Ooyen[1], Vivian Kwee and Roel Nusse

Department of Molecular Biology, Antoni van Leeuwenhoek Huis, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

[1]Present address: Gist Brocades, Postbus 1, 2600 MA Delft, The Netherlands

Communicated by P.Borst

The mouse mammary tumor virus can induce mammary tumors in mice by proviral activation of an evolutionarily conserved cellular oncogene called *int*-1. Here we present the nucleotide sequence of the human homologue of *int*-1, and compare it with the mouse gene. Like the mouse gene, the human homologue contains a reading frame of 370 amino acids, with only four substitutions. The amino acid changes are all in the hydrophobic leader domain of the *int*-1 encoded protein, and do not significantly alter its hydropathic index. The conservation between the mouse and the human *int*-1 genes is not restricted to exons; extensive parts of the introns are also homologous. Thus, *int*-1 ranks among the most conserved genes known, a property shared with other oncogenes.
Key words: conservation/human oncogene/nucleotide sequence

Introduction

Mammary tumors in mice infected with the mouse mammary tumor virus (MMTV) offer an experimental route to the identification of cellular oncogenes: by molecular cloning of host-cell DNA adjacent to integrated MMTV proviral elements (Varmus, 1984). This approach has led to the discovery of two different genes, *int*-1 and *int*-2, that are often trancriptionally activated as a consequence of nearby proviral insertions in mammary tumors (Nusse and Varmus, 1982; Nusse *et al.*, 1984; Peters *et al.*, 1983). The proviral insertions at the *int*-1 locus strongly implicate expression of an intact gene product in tumorigenesis; many insertions were found in the transcriptional unit of the gene, but the protein-encoding domain is always left intact (Van Ooyen and Nusse, 1984). This protein is, as deduced from the nucleotide sequence, 370 amino acids long, and does not resemble any known gene product.

The *int*-1 gene is, as many cellular oncogenes are, conserved in evolution, with homologous sequences in organisms ranging from *Drosophila* to man (Nusse *et al.*, 1984). We have cloned the human homologue of *int*-1, and assigned it to chromosome 12 (Van 't Veer *et al.*, 1984). The homology between the mouse and the human *int*-1 DNA was illustrated by a heteroduplex analysis. In this paper, we present the complete nucleotide sequence of the human *int*-1 gene and compare it with that of the mouse gene. Extensive homologies are found, both in the protein-encoding domain and in some intron areas.

Results and Discussion

Sequence analysis of the human int-1 gene

The human *int*-1 gene has been cloned as part of a 13.2-kb *Eco*RI fragment from a bacteriophage library of human placental DNA. The approximate position of the gene on this DNA fragment was first determined by the heteroduplex analysis presented before (Van 't Veer *et al.*, 1984), and subsequently by restriction enzyme sites which were conserved between the human and mouse genes. The precise position of the gene was obtained from the nucleotide sequence of the homologous area. The sequencing strategy and the resulting structure of the human *int*-1 gene (see below) is shown in Figure 1. Arrows indicate the direction and extent of sequencing. All protein-encoding sequences and most of the non-coding parts were determined from both strands.

Comparison of the human and mouse int-1 sequences and derivation of the structure of the human gene

In Figure 2 we have aligned the nucleotide sequence of the mouse and human *int*-1 genes. The following strategy for lining up both sequences was used. First, the human gene was compared with the mouse sequence in blocks of 50 nucleotides, with the aid of a computer program. Subsequently, regions of homology of 60% or more were lined up according to previously published rules, to obtain maximal homology (Van Ooyen *et al.*, 1979). Deletions and insertions were introducd in the sequence of the mouse gene if necessary. Figure 2 presents the result of this comparison; the human *int*-1 gene is given as a continuous sequence.

The structure of the human gene was derived by comparison with the previously established exon-intron structure of the mouse homologue (Van Ooyen and Nusse, 1984). This structure is based on extensive S1 mapping data and on unpublished cDNA cloning experiments (H.E.Varmus, personal communication). The mouse gene consists of four exons, the last of which contains a long untranslated trailer region and a polyadenylation signal. The first exon that has been detected contains the translational start signal and a non-translated leader. This exon is preceded by a TATA box which is conserved, leading us to conclude tentatively that this signal indicates the authentic start of the *int*-1 gene, but it cannot be excluded that transcription starts further upstream. Possible upstream exons must be non-coding, because the ATG

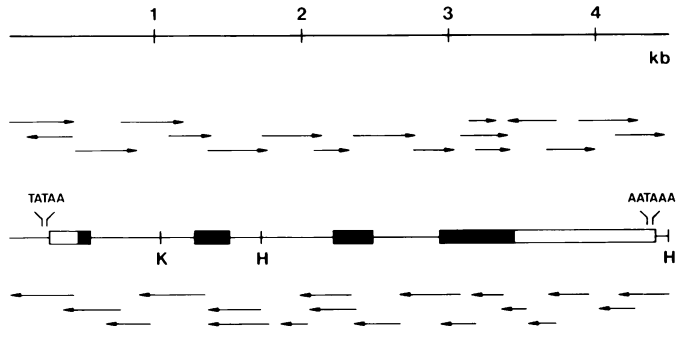
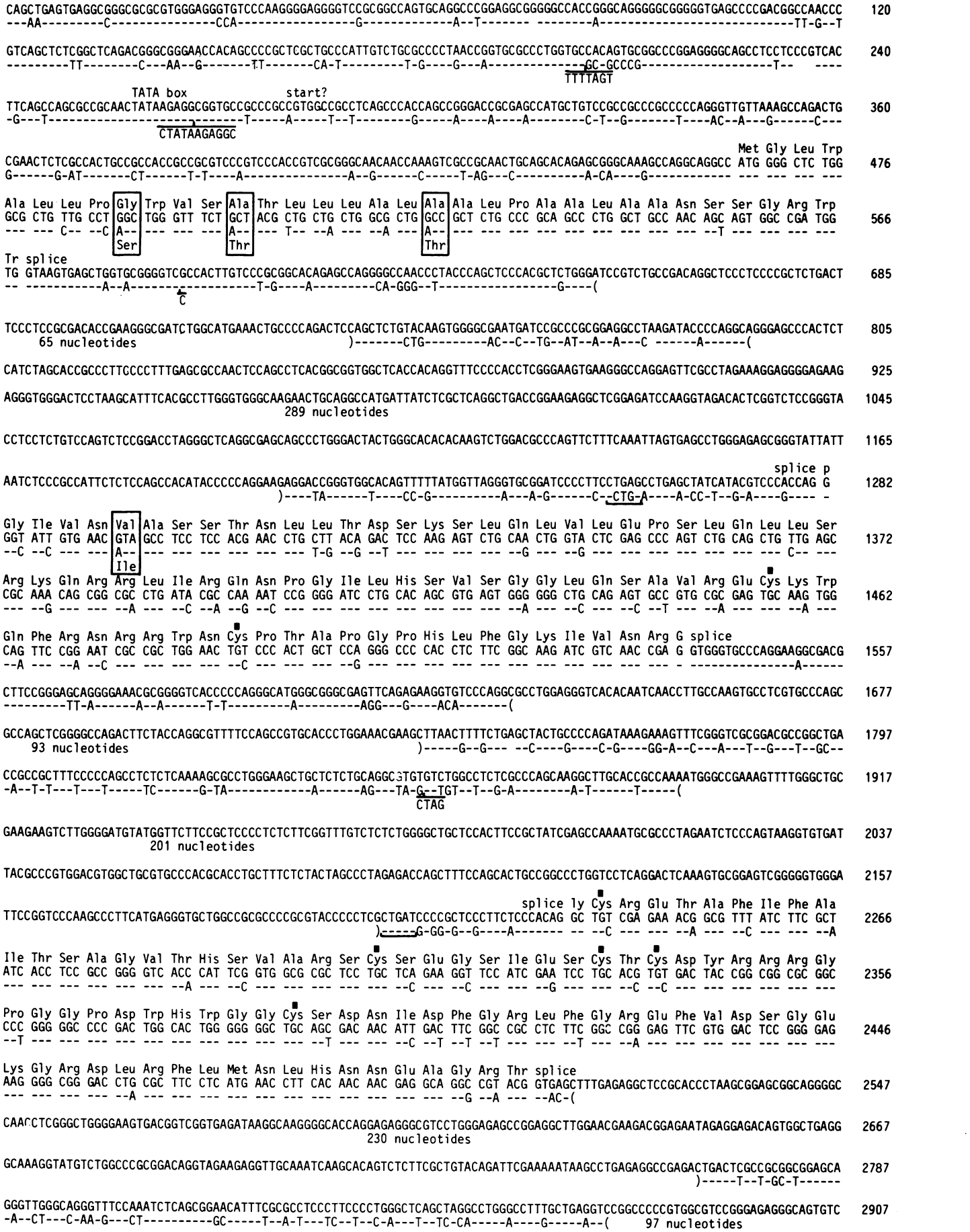


Fig. 1. Sequencing strategy and structure of the human *int*-1 gene. Arrows indicate direction and extent of reading of the sequenced fragments. Arrows in the upper part represent (+) strand sequences and in the lower part (−) strand sequences. The structure of the gene is derived from a comparison with the corresponding mouse gene (Figure 2); blocks represent exons; coding sequences are black; K: *Kpn*I; H: *Hind*III.



2906

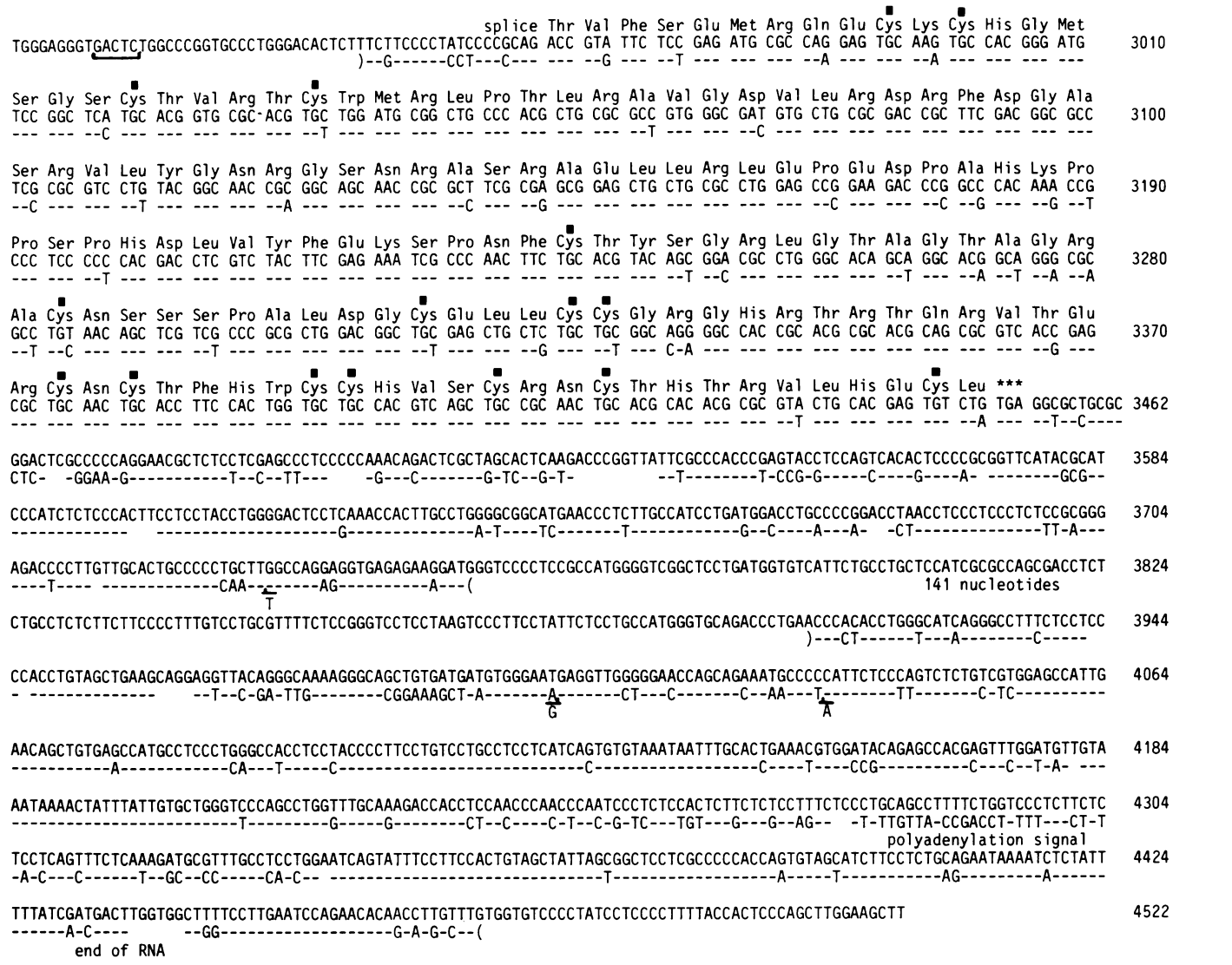


Fig. 2. Sequence of the human *int*-1 gene and comparison with the mouse homologue. The plus strand sequence of the human gene is shown as a continuum (top line). The mouse gene is aligned by rules explained in the text. Only regions of high homology are shown. Identical nucleotides are indicated by dashes; extra nucleotides are indicated in the mouse sequences; empty spots represent gaps. Regions of low homology are between brackets; the number of nucleotides in the corresponding area of the mouse sequence is shown. The amino acid sequence of the human gene, as well as splice sites, TATA box and polyadenylation signal are derived by comparison with the mouse gene. Amino acids which are different in both genes are boxed. Cysteines are indicated by dots. Variants of the conserved sequence CTGAC involved in lariat formation during splicing are underlined.

start codon is preceded by stop codons in all three reading frames (Van Ooyen and Nusse, 1984).

The comparison of the mouse with the human *int*-1 gene shows conservation of all splice sites and transcriptional-regulation sequences (TATA box and polyadenylation signal). The derived intron-exon structure of the human gene is shown diagramatically in Figure 1.

The human int-1 protein

The human *int*-1 protein, as deduced from the nucleotide sequence, is very similar to the mouse protein. Both proteins are 370 amino acids long, with only four differences (boxed in Figure 2). These substitutions are found near the amino terminus of the protein, which is markedly hydrophobic. A hydropathicity profile of the human *int*-1 protein is presented in Figure 3, showing a typical hydrophobic leader domain. The four amino acid differences from the mouse gene are within this leader, but do not influence the hydropathicity profile. Thus, if the hydrophobic leader is cleaved off during transport, the entire active domain of the *int*-1 protein is evolutionarily conserved between mouse and man. This extraordinary degree of sequence constraint indicates how essential the structure of the *int*-1 product is in determining its function.

Nevertheless, the normal function of *int*-1 remains enigmatic. The gene has been found to be expressed only in mouse mammary tumors bearing a proviral integration near the transcriptional unit, and not in any normal tissue tested so far. As reported before for the mouse *int*-1 protein, we have not found any similarity to other gene products (Van Ooyen and Nusse, 1984). A salient feature of the *int*-1 protein is the high content of cysteine residues, often in pairs, near the COOH terminus, as indicated by dots in Figure 2. Properties such as a hydrophobic leader and cysteine-rich domains have been found in receptor molecules, for example the epidermal growth factor, the low density lipoprotein and the insulin receptors, but the *int*-1 protein lacks the typical transmembrane domain of these receptors (Ullrich *et al.*, 1984, 1985; Yamamoto *et al.*, 1984; Ebina *et al.*, 1985). Alternative-

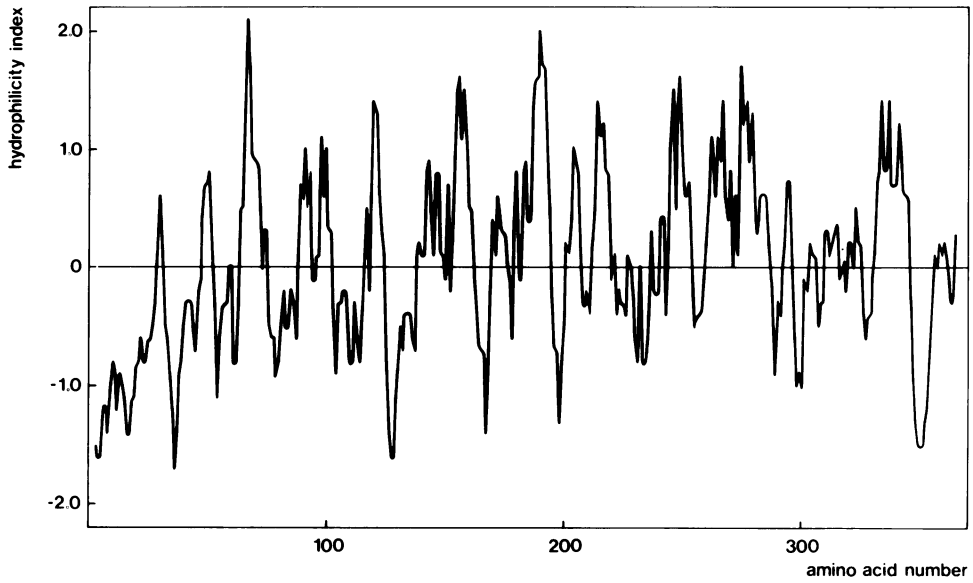


Fig. 3. Hydropathicity profile of *int*-1. Hydropathicity values for amino acids were taken from Hopp and Woods (1981). At every position the hexapeptide value was taken.

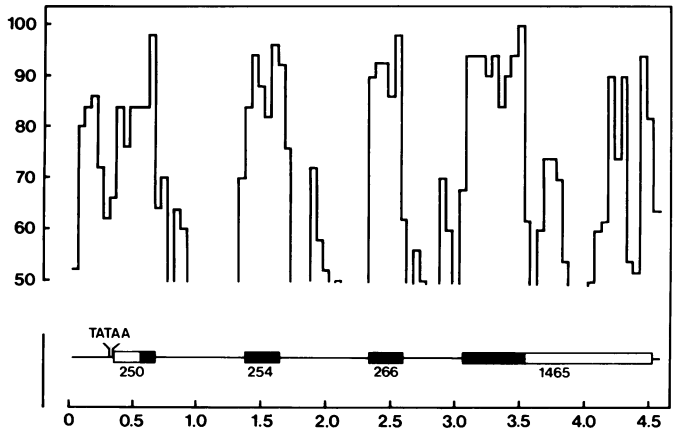


Fig. 4. Overall homology between human and mouse *int*-1. The human and mouse genes were compared in blocks of 50 nucleotides with the aid of a computer, without introduction of gaps in either of the sequences. Regions with >50% homology are shown. Exons are indicated by blocks, coding sequences are black.

ly, we speculate that the mature form of the *int*-1 protein is a growth factor. These molecules are often cysteine-rich and synthesized as precursors with hydrophobic leaders (cf. Feramisco *et al.*, 1985).

Conservation of non-coding sequences

The sequence comparison in Figure 2 shows that the homology between the mouse and the human gene is not restricted to the coding sequences. Besides homologies in the non-coding part of the mRNA sequence, extensive parts of the *int*-1 introns are highly conserved, most notably near the 5′ end of the first and the second intron, and near the 3′ end of the first intron. This homology extends far beyond what is generally found at intron boundaries (Breathnach and Chambon, 1981). A less pronounced, but significant homology is found in the center of all three introns. A diagram of the overall homology is shown in Figure 4.

The significance of these conservations is not clear, but explanations that come to mind are functions in regulation of gene expression or in splicing. A variant of the consensus sequence CTGAC involved in lariat formation during splicing (Keller and Noon, 1984) is found in all three introns (underlined in Figure 2) at 19–46 nucleotides from the 3′ splice site. These sequences are within conserved areas, with the exception of the TTGAC sequence in the third intron, which is in a region of low overall homology.

Conserved areas in introns have also pointed to the existence of elements regulating specific gene expression, or enhancers, in immunoglobulin genes (Emorine *et al.*, 1983; Gillies *et al.*, 1983; Banerji *et al.*, 1983; Queen and Baltimore, 1983). Other genes, globin and thymidine kinase, for example, also carry intragenic regulatory elements (Wright *et al.*, 1984; Merrill *et al.*, 1984). The large area of homology upstream from the first exon could also serve a function in regulating gene expression, as it may contain the *int*-1 promoter.

None of the homologous areas outside the *int*-1 exons can encode proteins of substantial length since stop codons occur frequently.

Implications

The sequence comparison between the mouse and the human *int*-1 homologues presented here illustrates the evolutionary conservation that is characteristic for oncogenes (Bishop, 1983); and is even unprecedented within the oncogene family. Compared with those oncogenes for which both the murine and the human cellular homologues have been sequenced, the overall amino acid sequence homology of *int*-1 (99%) is higher than that of c-*myc* (93%, Bernard *et al.*, 1983), of c-*fos* (90%, Verma *et al.*, 1984) and of c-*mos* (75%, Blair *et al.*, 1984), of p53 (78%, Zakut-Hori *et al.*, 1985) and of c-Ki-*ras* (97%, George *et al.*, 1985).

The mouse *int*-1 gene can be activated by proviral insertions of MMTV, leading to some step in mammary carcinogenesis. Whether abnormal expression of the virtually identical human homologue can contribute to mammary tumorigenesis in man, in which no manifest replication of MMTV-like viruses has been observed, remains to be seen. Rather than by proviral insertion, activation of the gene might well occur by gene amplification, the hallmarks of which – double minute chromosomes – are often observed in human mammary tumors (Barker and Hsu, 1979; Gebhardt *et al.*, 1984).

Materials and methods

DNA sequence analysis

DNA sequencing was carried out according to the method of Maxam and Gilbert (1980). Labeling of 5′ ends was carried out by treatment of restriction sites with calf intestine alkaline phosphatase (Boehringer) and phosphorylation with [γ-^{32}P]ATP (Amersham) and T4 polynucleotide kinase (Boehringer). Restriction fragments were labeled at 3′ ends with the large fragment of DNA polymerase I (New England Biolabs) and all four deoxynucleotides. Unlabeled nucleotides were at 20 μM and the labeled [α-^{32}P]dNTP at 1.5 μM.

Acknowledgements

We thank Mrs N.Nuland for secretarial assistance, especially in presenting the nucleotide sequence data, Piet Borst, David Greaves and Betsy Matthews for comments on the manuscript, and Harold Varmus for communicating unpublished results on cDNA cloning experiments. AvO and VK were supported by a grant from the Netherlands Cancer Society 'Koningin Wilhelmina Fonds'.

References

Banerji,J., Olson,L. and Schaffner,W. (1983) *Cell*, **33**, 729-740.
Barker,P.E. and Hsu,T.C. (1979) *J. Natl. Cancer Inst.*, **62**, 257-261.
Bernard,O., Cory,S., Gerondakis,S., Webb,E. and Adams,J.M. (1983) *EMBO J.*, **2**, 2375-2383.
Bishop,J.M. (1983) *Annu. Rev. Biochem.*, **52**, 301-354.
Blair,D.G., Wood,T.G., Woodworth,A.M., McGeady,M.L., Oskarsson,M.K., Propst,F., Tainsky,M.K., Cooper,C.S., Watson,R., Baroudy,B.M. and Vande Woude,G.F. (1984) *Cancer Cells*, **2**, 281-291.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.
Ebina,Y., Buis,L., Jarnagin,K., Edery,M., Graf,L., Clauser,E., Ou,J.H., Masiarz,F., Kan,Y.W., Goldfine,I.D., Roth,R.A. and Rutter,W.J. (1985) *Cell*, **40**, 747-758.
Emorine,L., Kuehl,M., Weir,L., Leder,Ph. and Max,E.E. (1983) *Nature*, **304**, 447-449.
Feramisco,J., Ozanne,B. and Stiles,C. (1985) *Cancer Cells, Vol. 3, Growth Factors and Transformation*, published by Cold Spring Harbor Laboratory Press, NY.
Gebhart,E., Brüderlein,S., Tulusan,A.H., Maillot,K.V. and Birkman,J. (1984) *Int. J. Cancer*, **34**, 369-373.
George,D.L., Scott,A.F., Trusko,S., Glick,B., Ford,E. and Dorney,D.J. (1985) *EMBO J.*, **4**, 1199-1203.
Gillies,S.D., Morrison,S.L., Oi,V.T. and Tonegawa,S. (1983) *Cell*, **33**, 717-728.
Hopp,T.P. and Woods,K.R. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 3824-3828.
Keller,E.B. and Noon,W.A. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 7417-7420.
Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499-560.
Merril,G.F., Hauschka,S.D. and McKnight,S.L. (1984) *Mol. Cell. Biol.*, **4**, 1777-1784.
Nusse,R. and Varmus,H.E. (1982) *Cell*, **31**, 99-109.
Nusse,R., Van Ooyen,A., Cox,D., Fung,Y.K.T. and Varmus,H.E. (1984) *Nature*, **307**, 131-136.
Peters,G., Brookes,S., Smith,R. and Dickson,C. (1983) *Cell*, **33**, 369-377.
Queen,C. and Baltimore,D. (1983) *Cell*, **33**, 741-748.
Ullrich,A., Coussens,L., Hayflick,J.S., Dull,T.J., Gray,A., Tam,A.W., Lee,J., Yarden,Y., Libermann,T.A., Schlessinger,J., Downward,J., Mayes,E.L.V., Whittle,N., Waterfield,M.D. and Seeburg,P.H. (1984) *Nature*, **309**, 418-425.
Ullrich,A., Bell,J.R., Chen,E.Y., Herrera,R., Petruzzelli,L.M., Dull,T.J., Gray, A., Coussens,L., Liao,Y.-C., Tsubokawa,M., Mason,A., Seeburg,P.H., Grunfeld,C., Rosen,O.M. and Ramachandran,J. (1985) *Nature*, **313**, 756-761.
Van Ooyen,A., Van den Berg,J., Mantei,N. and Weissmann,C. (1979) *Science (Wash.)*, **206**, 337-344.
Van Ooyen,A. and Nusse,R. (1984) *Cell*, **39**, 233-240.
Van 't Veer,L.J., Van Kessel,G., Van Heerikhuizen,H., Van Ooyen,A. and Nusse,R. (1984) *Mol. Cell. Biol.*, **4**, 2532-2534.
Varmus,H.E. (1984) *Annu. Rev. Genet.*, **18**, 553-612.
Verma,I.M., Curran,T., Müller,R., Van Straaten,F., MacConnel,W.P., Miller, A.D. and Van Beveren,C. (1984) *Cancer Cells*, **2**, 309-321.
Wright,S., Rosenthal,A., Flavell,R. and Grosveld,F. (1984) *Cell*, **38**, 265-273.
Yamamoto,T., Davis,C.G., Brown,M.S., Schneider,W.J., Casey,M.L., Goldstein,J.L. and Russell,D.W. (1984) *Cell*, **39**, 27-38.
Zakut-Hori,R., Bienz-Tadmor,B., Givol,D. and Oren,M. (1985) *EMBO J.*, **4**, 1251-1255.

Received on 22 July 1985