The EMBO Journal vol.4 no.13A pp.3483–3488, 1985

Nucleotide sequence of the L1 ribosomal protein gene of *Xenopus laevis*: remarkable sequence homology among introns

Fabrizio Loreni, Ida Ruberti, Irene Bozzoni, Paola Pierandrei-Amaldi[1] and Francesco Amaldi[2]

Centro Acidi Nucleici, CNR, Dipartimento di Genetica e Biologia Molecolare, I Università di Roma, 'La Sapienza', 00185 Roma, [1]Istituto di Biologia Cellulare, CNR, via Romagnosi 18/A, 00196 Roma, and [2]Dipartimento di Biologia, II Università di Roma 'Tor Vergata', via Orazio Raimondo, 00173 Roma, Italy

Communicated by F.Amaldi

Ribosomal protein L1 is encoded by two genes in *Xenopus laevis*. The comparison of two cDNA sequences shows that the two L1 gene copies (L1a and L1b) have diverged in many silent sites and very few substitution sites; moreover a small duplication occurred at the very end of the coding region of the L1b gene which thus codes for a product five amino acids longer than that coded by L1a. Quantitatively the divergence between the two L1 genes confirms that a whole genome duplication took place in *Xenopus laevis* ~30 million years ago. A genomic fragment containing one of the two L1 gene copies (L1a), with its nine introns and flanking regions, has been completely sequenced. The 5′ end of this gene has been mapped within a 20-pyridimine stretch as already found for other vertebrate ribosomal protein genes. Four of the nine introns have a 60-nucleotide sequence with 80% homology; within this region some boxes, one of which is 16 nucleotides long, are 100% homologous among the four introns. This feature of L1a gene introns is interesting since we have previously shown that the activity of this gene is regulated at a post-transcriptional level and it involves the block of the normal splicing of some intron sequences.

Key words: DNA sequence/intron/ribosomal protein L1/*Xenopus laevis*

Introduction

Eukaryotic ribosome biosynthesis is a complex process which involves the co-regulated expression of many genes coding for its structural components, the rRNA and the ribosmal proteins (r-proteins). In recent years several r-protein mRNAs and genes from various eukaryotic systems have been cloned in order to elucidate the molecular mechanisms involved in their coordinate expression (for a review, see Fried and Warner, 1984).

We have previously reported the construction, isolation and characterization of cDNA clones specific for six different r-proteins of *Xenopus laevis* (Pierandrei-Amaldi and Beccari, 1980; Bozzoni *et al.*, 1981; Amaldi *et al.*, 1982) and we have used them to study the expression of r-protein genes in *Xenopus* oocyte and embryo development (Pierandrei-Amaldi *et al.*, 1982, 1985). One of these cDNA clones, pXom102, contains the 3′ portion of the mRNA for r-protein L1. We have shown that the corresponding gene is present in two copies per haploid genome in *X. laevis* (Bozzoni *et al.*, 1981). One of the two L1 gene copies was isolated from a *X. laevis* genomic library and its structure analyzed at the level of restriction map and exon-intron organization (Bozzoni *et al.*, 1982): it is ~6 kb long and consists of 10 exons, of sizes ranging from ~60 to 200 bp, and nine introns. To study the mechanisms involved in its regulation, the cloned L1 gene has been microinjected into *Xenopus* oocytes. These experiments have shown that a specific block of processing of the L1 transcript, resulting in the retention of two of the nine introns, is responsible for its regulation (Bozzoni *et al.*, 1984). Here we report the nucleotide sequence of the entire L1 gene, including the nine introns which have been searched for structural features which might be responsible for the described regulation at the processing level. We also report and compare the sequences of two newly isolated cDNA clones which represent the transcription products of the two L1 gene copies.

Results and Discussion

Isolation, analysis and comparison of cDNA clones for L1a and L1b r-proteins

Since the amino acid sequence of r-protein L1 is unknown, the identification of exon and intron positions along the gene requires the comparison with the nucleotide sequence of a full-length cDNA for the same r-protein. On the other hand the cDNA clone specific for *X. laevis* L1 protein which we previously isolated (Bozzoni *et al.*, 1981) and sequenced (Amaldi *et al.*, 1982) corresponds to only about one third of the mRNA at its 5′ end. For this reason we have now screened, using our cDNA clone as probe, another cDNA library constructed in λgt10 with poly(A)$^+$ RNA from *Xenopus* oocytes by D.Melton and kindly made available to us. Several clones have been isolated which fall into two classes, by restriction map analysis, as we might have expected on the basis of the presence of two L1 gene copies in the *X. laevis* genome. Two clones, one for each class, have been sequenced.

The strategies used to sequence these two cDNA inserts are shown in Figure 1a and a′, and the sequences are presented and compared in Figure 2. One of the two cDNAs (L1a) falls short of a full length mRNA as it almost reaches the 5′ end of the gene (see below), while the other (L1b) lacks ~50 nucleotides belonging to the first exon. The comparison of the two cDNA sequences reveals that the two L1 gene copies have somewhat diverged since the gene duplication occurred. Most of the nucleotide substitutions are observed at silent sites, thus leaving the protein primary structure unchanged. The few cases where replacement mutations have occurred led to conservative changes which result in the substitution of an amino acid with another of similar properties. This divergence between the two L1 gene copies is quantitatively very similar to that found between the two copies of the vitellogenin A gene and between the two copies of the vitellogenin B gene in this same species (Germond *et al.*, 1984), and agrees with the notion that a whole genome duplication took place in *Xenopus* ~30 million years ago (Bisbee *et al.*, 1977). It is interesting to notice that the 3′-untranslated regions, ~60 nucleotide long, are almost identical in the two cDNAs, suggesting that this part of the RNAs, although not coding, has some important function.

The two cDNAs differ also in the total length of the coding

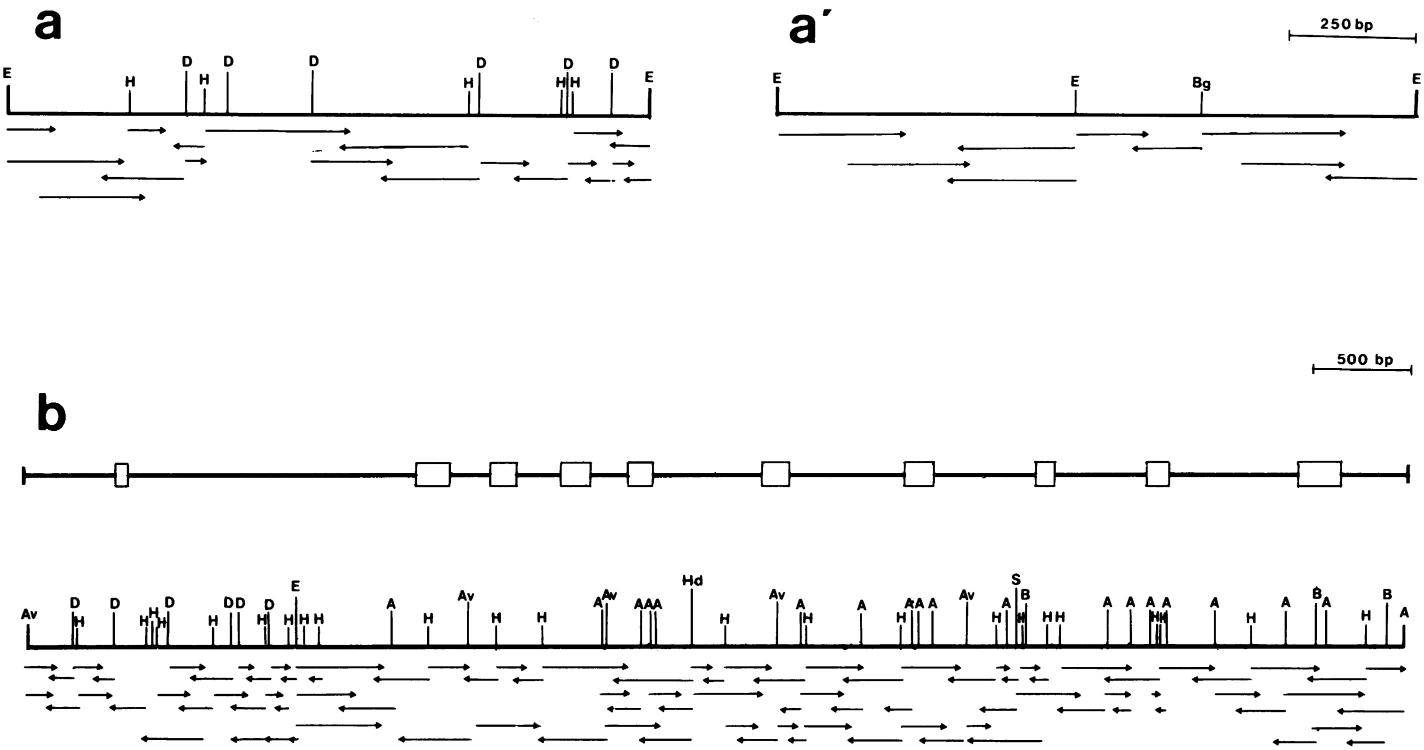

Fig. 1. Sequencing strategies for L1a cDNA (**a**), L1b cDNA (**a'**), and for the *X. laevis* genomic fragment containing the L1a gene (**b**); for this, the exon-intron organization is also shown. Only the restriction sites used for sequencing are indicated in the maps. Abbreviations: E, *Eco*RI; H, *Hinf*I; D, *Dde*I; Bg, *Bgl*II; Av, *Ava*II; A, *Alu*I; Hd, *Hind*III; S, *Sal*I; B, *Bsp*RI. Arrows indicate the direction and extent of sequence analysis. Arrows starting at positions not corresponding to any restriction site indicate sequencing of subcloned fragments obtained by *Bal*31 digestion.

region. In fact the L1b mRNA (cDNA) has a 15-nucleotide insertion, with respect to L1a, at the very end of the coding region before the TAA stop codon. A closer look at this 15-nucleotide sequence suggests that it might have originated by duplication of the 15-nucleotide sequence just preceding it (also present in L1a mRNA), followed by some divergence. On the other hand when we analyze r-proteins by two-dimensional gel electrophoresis, although the L1 protein appears generally as a single spot (Pierandrei-Amaldi and Beccari, 1980), sometimes a slight separation of more than one component can be seen. The resolution of the L1 spot into three components can be obtained by changing the acrylamide-bisacrylamide ratio of the second dimension SDS gel. Figure 3 shows the resolution of the three L1 components by one-dimensional SDS gel electrophoresis of r-proteins from the large subunit. The first two of these three components are indistinguishable by peptide mapping after digestion with *Staphylococcus aureus* V8 protease, and represent the products coded by the two L1 gene copies (as indicated by hybrid-selected translation experiments; not shown). The small difference in mol. wt. between the two L1 proteins is probably due to the short duplication, described above, at the end of the coding region of L1b cDNA. Both genes appear to be active (L1a more than L1b) in the *X. laevis* oocytes, as shown by the isolation of L1a and L1b sequences from the cDNA bank and by the observation of the two resolved L1 spots in the electrophoretic patterns of r-proteins prepared from oocytes. The third component now resolved from the original L1 spot has a completely different peptide map and probably represents another r-protein (Lx in Figure 3).

In Figure 2 we have indicated as initiation codon the first ATG followed by the long open reading frame. The next ATG on the same reading frame is five codons downstream. If the first ATG is the functional initiation signal the protein product coded by the L1a gene contains 396 amino acids, has a mol. wt. of 44 920 daltons and, as expected, is basic in character: arginine, lysine and histidine together represent 25% of the residues while <6% are glutamic and aspartic acids. We do not consider L1b here as it is incomplete at the amino terminus.

Nucleotide sequence of the L1a gene

Figure 1b shows the structural map of the portion of the *X. laevis* genomic fragment inserted in λXlrp14 (Bozzoni *et al.*, 1982), which contains one of the two gene copies coding for r-protein L1 (gene L1a). Figure 1b diagrams the strategy and the restriction enzyme sites used to obtain the L1a gene sequence which is presented in Figure 4, including several hundred nucleotides of the flanking regions. The comparison of this sequence with those of the two cDNAs described above allowed the precise identification of the position of exons and introns in this genomic fragment. The same comparison allowed the identification of the position of the cleavage/polyadenylation site at the 3' end of the gene; it has been found to be located 15 nucleotides downstream of the polyadenylation signal (AAUAAA) in both cDNAs, while in our previous L1 cDNA clone (corresponding to a L1b) it was located at 13 nucleotides from the AAUAAA. Similar 3' microheterogeneity has been described for other gene transcription products in particular for mouse r-protein L30 (Wiedemann and Perry, 1984).

The 5' end of the gene has been positioned with a few nucleotides of uncertainty by a primer extension experiment (Figure 5). In a previous paper (Bozzoni *et al.*, 1982) we localized it several hundred nucleotides downstream of the site now identified. We know now that this 5' end localization, which was obtained by S1 mapping, refers to an unrelated RNA originating from transcription of the other strand (unpublished data). The now identified 5' end of the L1a gene is located nine (plus or minus two) nucleotides upstream from the 5' terminus of the clon-

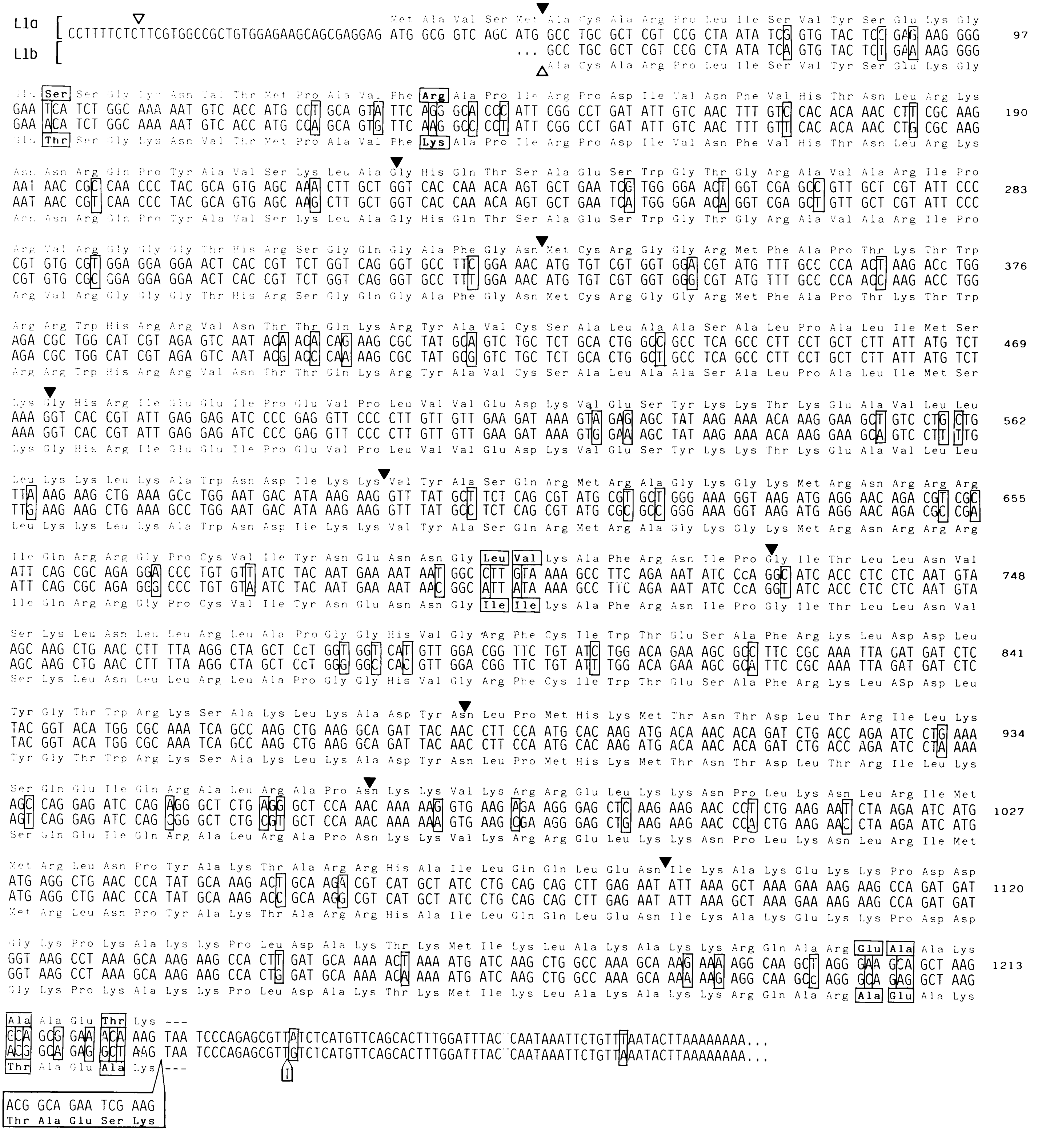

Fig. 2. Comparison of the nucleotide sequence of L1a and L1b cDNA and of the deduced amino acid sequences. Nucleotide and amino acid substitutions are boxed. Closed triangles indicate the positions of introns in the L1a gene. The open triangles indicate the 5′ ends of the cloned L1a and L1b cDNAs; the few nucleotides which complete the L1a sequence upstream have been deduced from the L1a gene sequence (Figure 4) and the primer extension experiment (Figure 5).

ed L1a cDNA; it falls within a 20 pyrimidine residue tract and is preceded upstream (−25) by a reasonable TATA-like sequence. The presence of a pyrimidine-rich 5′ end has been described in two mouse r-protein genes (Wiedemann and Perry, 1984; Dudov and Perry, 1984), not in the r-protein 49 gene of *Drosophila* (O'Connell and Rosbash, 1984) and only in some of the several yeast r-protein genes analyzed (for references see Teem *et al.*, 1984). On the other hand it has been noted that this type of 5′ end is common to several 'housekeeping' eukaryotic genes (Dudov and Perry, 1984).

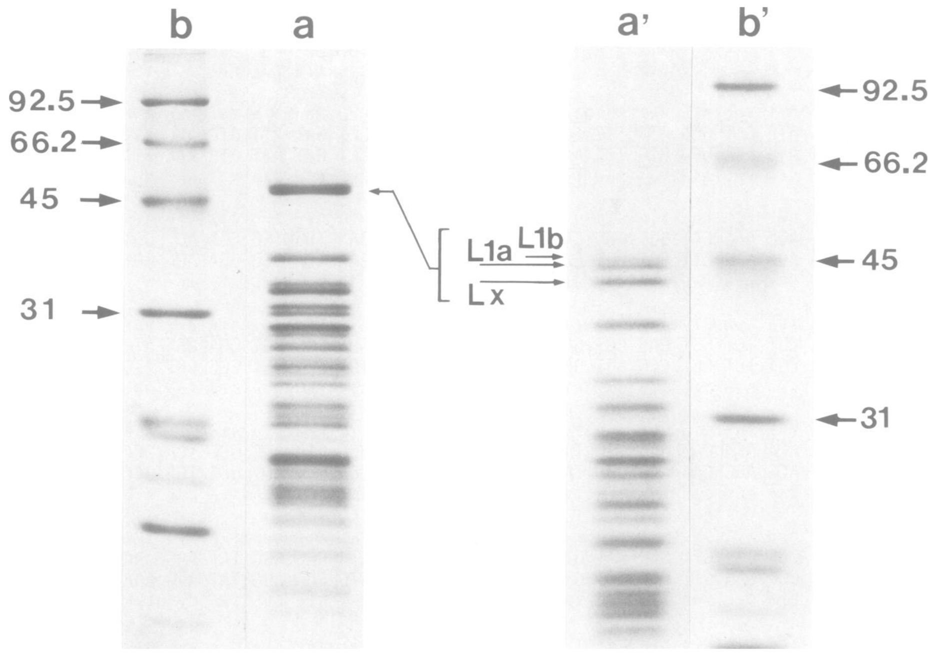

Fig. 3. Resolution of L1a and L1b r-proteins on SDS gel electrophoresis. Proteins extracted from *X. laevis* large ribosomal subunits were run on acrylamide gel electrophoresis (15% with SDS) at different acrylamide-bisacrylamide ratios: 35/1 (**a**); 140/1 (**a'**) together with size markers (**b** and **b'**), whose mol. wt. values are given in kd.

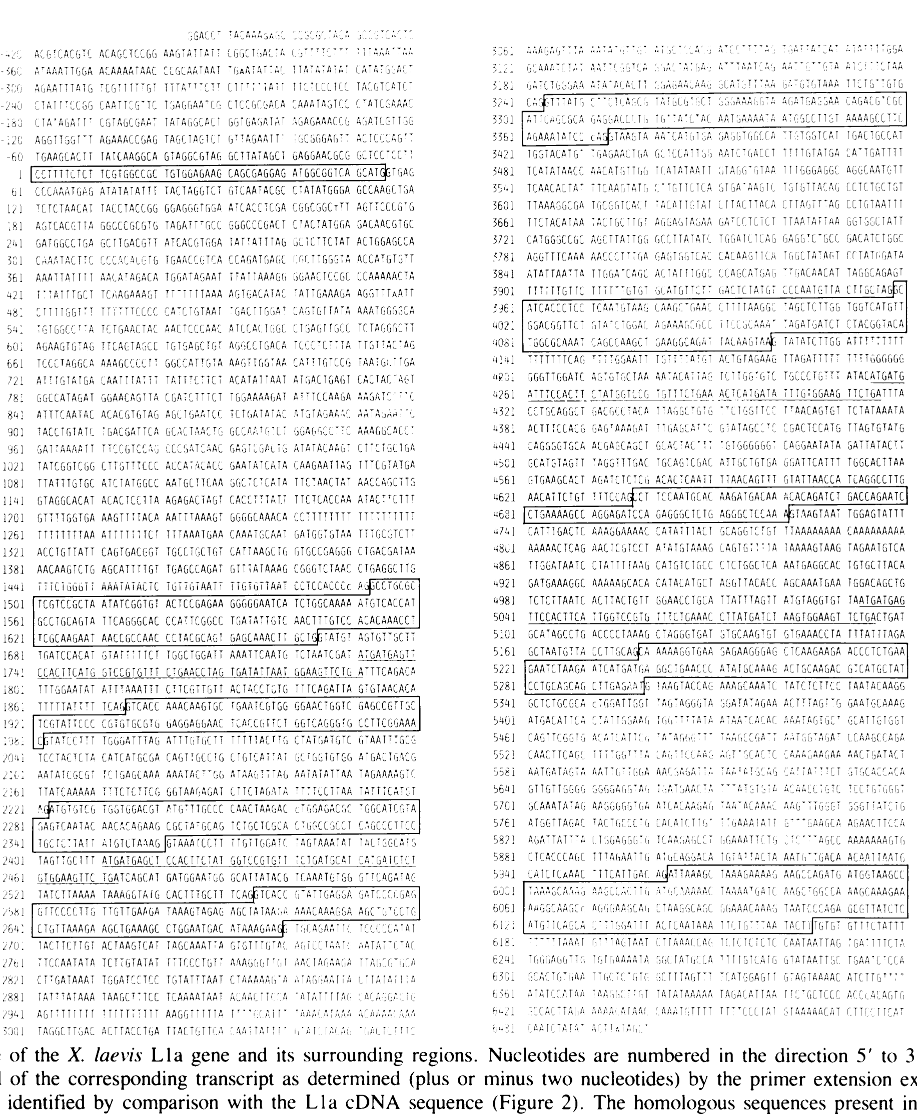

Fig. 4. Nucleotide sequence of the *X. laevis* L1a gene and its surrounding regions. Nucleotides are numbered in the direction 5' to 3', starting with number 1 at the position of the 5' end of the corresponding transcript as determined (plus or minus two nucleotides) by the primer extension experiment (Figure 5). The 10 exons (boxes) have been identified by comparison with the L1a cDNA sequence (Figure 2). The homologous sequences present in introns 2, 4, 7 and 8 are underlined. Lower-case letters indicate uncertainty in nucleotide identification.

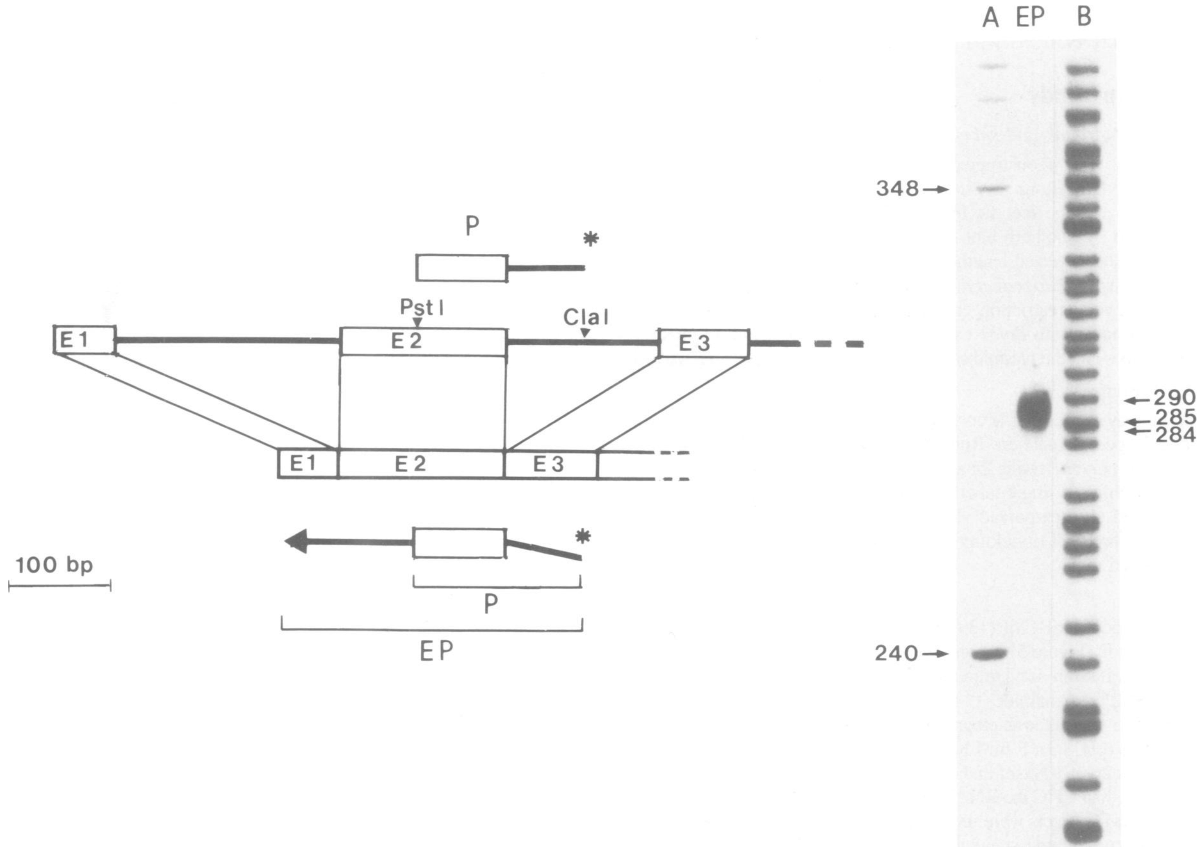

Fig. 5. Determination of the 5′ end by primer extension. A *Pst*I-*Cla*I fragment (139 bp) including part of the second exon and part of the second intron was isolated (P), end-labelled and used to prime cDNA synthesis from oocyte poly(A)$^+$ RNA as shown in the schematic diagram. The product (EP) was electrophoresed on acrylamide gel together with size markers (**A** and **B**).

```
...CATG:GTGAGCCCAA....   intron 1    ....TCCACCCCAG:GCCT...
                          1436 bp
...GCTG:GTATGTAGTG....   intron 2    ....TATTTTTCAG:GTCA...
                          210 bp
...AAAC:GTATCCTTTT....   intron 3    ....TTTCATGTAG:ATGT...
                          241 bp
...AAAG:GTAAATCCTT....   intron 4    ....TTGCTTTCAG:GTCA...
                          194 bp
...GAAG:GTGCAGAATT....   intron 5    ....TGTTGTGCAG:GTTT...
                          564 bp
...CCAG:GTAAGTAAAT....   intron 6    ....TACTTGCTAG:GCAT...
                          589 bp
...ACAA:GTAAGTATAT....   intron 7    ....TGTTTTCCAG:CCTT...
                          522 bp
...CAAA:GTAAGTAATT....   intron 8    ....TACCTTGCAG:CAAA...
                          457 bp
...GAAT:GTAAGTACCA....   intron 9    ....TCATTGACAG:ATTA...
                          663 bp

   A     A                               CCCCCC C      G
...CAG:GTGAGT....       consensus    ....TTTTTTNTAG:GT...
```

Fig. 6. Exon/intron and intron/exon junctions in the L1a gene are compared with the consensus sequence (Breathnach and Chambon, 1981).

Structural features of the introns of L1a gene

Figure 6 compares the exon-intron-exon junctions of the nine introns of the L1a gene. All the sequences obey the GT/AG rule and in general are in good agreement with the larger consensus observed by Breathnach and Chambon (1981).

We have previously shown that the expression of the L1 gene is regulated at a post-transcriptional level both *in vivo* during embryogenesis (Pierandrei-Amaldi *et al.*, 1985) and in oocytes injected with the cloned L1a gene (Bozzoni *et al.*, 1984). In particular we have been able to identify, in the oocyte system, a specific regulation of L1 r-protein synthesis. The observed regulation involves a splicing block which leads to the accumulation of a precursor RNA still containing the second and the third introns (Bozzoni *et al.*, 1984). With these notions in mind we have now performed a computer analysis of the sequence of the L1a gene, with particular attention paid to the introns. Although no peculiar structure common to the second and third introns has been found, the search for repeats in the gene has revealed a striking feature in four of the nine introns: as shown in Figure 7 a sequence of 60 nucleotides is present with 80% homology in introns 2, 4, 7 and 8 (no other significant repeated sequence has been found in this genomic fragment with the exception of several T-runs present mainly in the introns). A closer analysis of these sequences shows that within these 60 nucleotides there are five boxes, ranging in size from 4 to 16 bases, which are 100% homologous. No case has been described so far of such high sequence homology in different introns interrupting the same gene, besides sequences necessary for splicing at border junctions and branching sites. Sequence homology has been described only in that particular class of introns which autosplice or code for a maturase (for references, see Waring and Davies, 1984). These considerations suggest that the homologous sequence present in the four introns of L1a gene might be involved in some specific regulatory or structural function. Experiments are now in progress to test whether these sequences are important for the regulation of the L1 RNA maturation or whether other specific functions must be attributed to them; this possibility is suggested

```
INTRON 2   1728   GATATGATGAGTTCCACTTCA-TGGTCCGTGTTTCTGAACC-TAGTGATATTAATGGAAGTTCTGATTT   1794
INTRON 4   2408   TTTATGATGAGCTCCACTTCTATGGTCCGTGTTTCTGATGCATCATGATCTCTGTGGAAGTTCTGATCA   2476
INTRON 7   4252   TACATGATGATTTCCACTTCTATGGTCCGTGTTTCTGAAAC-TCATGATATTTGTGGAAGTTCTGATTT   4319
INTRON 8   5030   TTAATGATGAGTTCCACTTCATTGGTCCGTGTTTCTGAAACCTTATGATCTAAGTGGAAGTTCTGACTG   5098
```

Fig. 7. Comparison of the homologous sequences present in four introns of the *X. laevis* L1a gene. The positions of the first and last nucleotides of the four sequences are given according to the nucleotide numbering of Figure 4.

by the fact that the homologous sequence occurs in four introns only one of which is involved in processing regulation.

Materials and methods

Screening of the cDNA bank and subcloning

A full-length cDNA bank (constructed in λgt10 with poly(A)$^+$ RNA from *X. laevis* oocytes, by D.Melton) was screened using the clone pXom102 (Bozzoni *et al.*, 1981) as a probe specific for r-protein L1 sequences. The isolated clones were analyzed by Southern blot hybridization and those which appeared to contain inserts of the expected length were subcloned in pSP6. The genomic fragment clones in Xlrp14 (Bozzoni *et al.*, 1982), containing the entire L1a gene, was digested with several restriction enzymes. The fragments obtained (some of them were also treated with *Bal*31 exonuclease) were subcloned in pBR322. Both the lambda clones and the plasmid subclones were used for sequence analysis.

DNA sequence analysis

DNA sequencing was carried out according to Maxam and Gilbert (1980) with the addition of a T-specific reaction (Rubin and Schmid, 1980). Fragments were end-labelled with T4 polynucleotide kinase and strand separated or restricted with a second enzyme yielding a single labelled end. The chemical reaction products were electrophoresed on urea-polyacrylamide (30:1) gels 40 × 20 × 0.03 cm (a 20% and two or three 6%), yielding an average of 200 – 300 nucleotides of sequence per labelled end.

Primer extension

The restriction fragment *Pst*I-*Cla*I (139 bp) including parts of the second exon and of the second intron (Figure 5) was end-labelled with T4 polynucleotide kinase and the strands separated on acrylamide gel. This primer was annealed to 5 μg poly(A)$^+$ RNA in 80% formamide, 0.4 M NaCl and 40 mM MOPS (pH 6.5) at 44°C for 5 h. The mixture was ethanol precipitated and then resuspended in 20 μl of 50 mM Tris (pH 8.3), 5 mM $MgCl_2$, 40 mM KCl, 2 mM DTT, 1 mM of each dNTP, 20 units of RNasin and 13 units of AMV reverse transcriptase. After incubation for 2 h at 42°C the RNA was hydrolyzed with alkali. After phenol extraction the extended products were recovered by ethanol precipitation and electrophoresed in urea-acrylamide sequencing gel together with size markers.

Acknowledgements

We thank Mr A.Di Francesco and Mr G.Pisaneschi for excellent technical assistance. This work was supported by grants from 'Progetto Finalizzato Ingegneria Genetica e Basi Molecolari delle Malattie Ereditarie, CNR', from 'Istituto Pasteur – Fondazione Cenci Bolognetti' and from 'Ministero della Pubblica Istruzione'.

References

Amaldi,F., Beccari,E., Bozzoni,I., Luo,X.Z. and Pierandrei-Amaldi,P. (1982) *Gene*, **17**, 311-316.

Bisbee,C.A., Baker,M.A., Wilson,A.C., Hadji-Azimi,I. and Fischberg,M. (1977) *Science (Wash.)*, **195**, 785-787.

Bozzoni,I., Beccari,E., Luo,X.Z., Amaldi,F., Pierandrei-Amaldi,P. and Campioni,N. (1981) *Nucleic Acids Res.*, **9**, 1069-1086.

Bozzoni,I., Tognoni,A., Pierandrei-Amaldi,P., Beccari,E., Buongiorno-Nardeli,M. and Amaldi,F. (1982) *J. Mol. Biol.*, **161**, 357-371.

Bozzoni,I., Fragapane,P., Annesi,F., Pierandrei-Amaldi,P., Amaldi,F. and Beccari,E. (1984) *J. Mol. Biol.*, **180**, 987-1005.

Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.

Dudov,K.P. and Perry,R.P. (1984) *Cell*, **37**, 457-468.

Fried,H.M. and Warner,J.R (1984) in Stein,G.S. and Stein,J.L. (eds.), *Recombinant DNA and Cell Proliferation*, Academic Press, NY, pp. 169-192.

Germond,J.E., Walker,P., ten Heggeler,B., Brown-Luedi,M., de Bony,E. and Wahli,W. (1984) *Nucleic Acids Res.*, **12**, 8595-8609.

Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499-560.

O'Connell,P. and Rosbasch,M. (1984) *Nucleic Acids Res.*, **12**, 5495-5513.

Pierandrei-Amaldi,P. and Beccari,E. (1980) *Eur. J. Biochem.*, **106**, 603-611.

Pierandrei-Amaldi,P., Beccari,E., Bozzoni,I. and Amaldi,F. (1985) *Cell*, **42**, 317-323.

Pierandrei-Amaldi,P., Campioni,N., Beccari,E., Bozzoni,I. and Amaldi,F. (1982) *Cell*, **30**, 163-171.

Rubin,C.M. and Schmid,C.W. (1980) *Nucleic Acids Res.*, **8**, 4613-4619.

Teem,J.L., Abovich,N., Kaufer,N.F., Schwindlinger,W.F., Warner,J.R., Levy,A., Woolford,J., Leer,R.J., van Raamsdonk-Duin,M.M.C., Mager,W.H., Planta,R.J., Schultz,L., Friesen,J.D., Fried,H. and Rosbash,M. (1984) *Nucleic Acids Res.*, **12**, 8295-8312.

Waring,R.B. and Davies,R.W. (1984) *Gene*, **28**, 277-291.

Wiedemann,L.M. and Perry,R.P. (1984) *Mol. Cell. Biol.*, **4**, 2518-2528.

Received on 13 August 1985; revised on 2 October 1985