The EMBO Journal vol.9 no.12 pp.4073-4081, 1990

Gene structure of nuclear lamin LIII of *Xenopus laevis*; a model for the evolution of IF proteins from a lamin-like ancestor

Volker Döring[1] and Reimer Stick

Universität München, Institut für Genetik und Mikrobiologie, Maria-Ward-Strasse 1a, D-8000 München 19, FRG

[1]Present address: Max-Planck-Institut für Biochemie, D-8033 Martinsried bei München, FRG

Communicated by H. Jäckle

The lamin LIII gene of *Xenopus laevis* has been characterized. The gene is duplicated in the *Xenopus* genome. The transcribed region spreads over 22 kb of genomic DNA encoding 12 exons. Two alternatively spliced mRNAs are observed which encode LIII isoforms that differ only by the 12 C-terminal amino acids which, however, both contain the CaaX motif known to be the target of post-translational modifications. The intron pattern of the lamin LIII gene is strikingly similar to that of an invertebrate intermediate filament (IF) gene over the entire protein coding sequence. The similarity in gene structure is restricted to the rod domain when compared with vertebrate types I–III IF genes. Our data suggest a model of how IF proteins evolved from a lamin-like ancestor by deletion of two signal sequences; the nuclear localization signal and the C-terminal *ras*-related CaaX motif. The data rule out the previously proposed hypothesis that IF proteins evolved from an intronless ancestor with an early divergence of neuronal and non-neuronal IF proteins. Together with the data presented in the accompanying paper by Dodemond *et al.* it can be concluded that the tail domains of lamins and invertebrate IF proteins, but not those of vertebrate IF proteins, are homologous. Thus, the different vertebrate IF proteins probably evolved by combination of the central rod domain with different tail domains by exon shuffling.

Key words: evolution/gene structure/intermediate filament proteins/lamins/*Xenopus*

Introduction

The nuclear lamina is a karyoskeletal structure. It lines the nucleoplasmic side of the nuclear membrane as a fibrous layer (for review see Gerace and Burke, 1988). A lamina structure has been demonstrated in a wide variety of organisms and tissues ranging from protozoa to vertebrates (Pappas, 1956; Fawcett, 1966) and it seems to be a universal feature of eukaryotes. The major structural proteins of the nuclear lamina are the nuclear lamins. In invertebrates [molluscs and arthropods (Dessev and Goldman, 1990; Gruenbaum *et al.*, 1988)] only one lamin polypeptide has been defined. Vertebrates, in contrast, express a variety of lamin polypeptides (for review see Krohne and Benavente, 1986). Indirect evidence indicates also the presence of lamins in yeast (Georgatos *et al.*, 1989). The vertebrate lamins can be grouped into type A and type B lamins (Lehner *et al.*, 1986; Krohne *et al.*, 1987; Peter *et al.*, 1989; Vorburger *et al.*, 1989b). While type B lamins seem to be constitutively expressed, expression of type A lamins is highly regulated during development and cell differentiation (Benavente *et al.*, 1985; Stick and Hausen, 1985; Lehner *et al.*, 1987; Röber *et al.*, 1989 and citations therein). The functional significance of this regulation is not yet understood.

cDNA sequencing as well as structural analysis made clear that lamins and intermediate filament proteins show remarkable similarities (McKeon *et al.*, 1986; Fisher *et al.*, 1986; Aebi *et al.*, (1986). They show structural features previously defined as diagnostic for members of the IF protein family, i.e. a tripartite domain structure with a central rod domain subdivided into four (in lamins three) α-helical coils characterized by a heptad repeat of hydrophobic amino acids (Steinert and Roop, 1988). This sequence principle is responsible for the coiled-coil forming ability of IF proteins. Noticeable sequence similarity between IFs and nuclear lamins, however, is seen only at the ends of the rod domains, while along most of the rod domain conservation of the heptad sequence principle rather than actual sequences is found. IF proteins of invertebrates show in some aspects a closer relationship to lamins than their vertebrate counterparts (Weber *et al.*, 1988, 1989). Both nuclear lamins and IF proteins from molluscs and nematodes have an extra six heptads in coil segment 1b. (Weber *et al.*, 1988, 1989). This segment is not found in vertebrate IF proteins. Moreover, sequence comparison between invertebrate IF proteins and nuclear lamins reveals a moderate amino acid sequence similarity in their tail domains, while vertebrate IF proteins totally diverge in this domain (Weber *et al.*, 1988, 1989).

Nuclear lamins possess some characteristics that clearly distinguish them from IF proteins. They possess a nuclear localization signal that directs the lamins to the nuclear compartment (Loewinger and McKeon, 1988), and a C-terminal sequence motif CaaX, which is also found in yeast mating factors and in *ras* proteins (Hancock *et al.*, 1989). This motif serves as a recognition signal for the post-translational isoprenylation at the C-terminal cysteine residue as well as for proteolytic processing (Vorburger *et al.*, 1989a). The isoprenylation, which persists in B-type lamins, is involved in membrane association of these lamins (Gerace and Blobel, 1980; Stick *et al.*, 1988; Holtz *et al.*, 1989; Krohne *et al.*, 1989). Both of these signal sequences are absent from IF proteins. Furthermore, the nuclear lamina is disassembled and reassembled during mitosis parallel to a reversible hyperphosphorylation of the lamin polypeptides (Gerace and Blobel, 1980).

The vertebrate IF proteins have been divided into four major classes: the acidic and basic epidermal keratins form types I and II respectively, vimentin, desmin, the glial fibrillar acidic protein and peripherin form type III, and the neurofilaments form type IV (for review see Osborn and Weber, 1986; Steinert and Roop, 1988). The gene structure of members of all four vertebrate IF classes has been

analysed. IF genes of types I–III show striking similarity in the exon/intron pattern in the central rod domain. Five out of eight introns found in these genes show identical positions with respect to the common domain structure. In contrast to the central rod domain, intron positions in the tail domains show no obvious similarity between vertebrate IF genes of different types. The neurofilaments totally diverge in their gene structure. This has been interpreted in different ways. On the one hand it has been suggested that neurofilament genes lost previously present introns by a retrotransposition event (Lewis and Cowan, 1986) and acquired new introns after divergence. An alternative hypothesis assumes that the ancestral IF gene was not interrupted by introns but that introns were inserted into the separate types I–III and type IV progenitors after this divergence (Steinert and Roop, 1988).

The common structure of cytoplasmic IF proteins and nuclear lamins as well as the sequence similarities found between these two classes of proteins might either be based on common ancestry or could be explained by protein design constraints leading to convergent evolution of both types of filament forming proteins. The observed similarities have been interpreted in favour of an evolutionary relationship. However, since convergent evolution cannot be ruled out, a rigorous proof of this hypothesis cannot be based on amino acid sequence comparison alone but relies on the analysis of the gene structure of nuclear lamins. We have analysed the genomic structure of a *Xenopus laevis* lamin (lamin LIII) and have compared this with the structure of the vertebrate IF genes as well as with the gene structure of an invertebrate IF protein [Dodemont *et al.* (1990) see accompanying paper]. Our results show that there exists striking conservation of intron positions between the nuclear lamin gene and the IF genes in the central rod domain strongly suggesting a common ancestry of lamins and IF proteins. Moreover, a remarkable conservation in the gene structure of the tail domains of the invertebrate IF protein and lamin LIII shows that the tail domain of the lamin protein is a homologue of the tail domain of the invertebrate IF protein but not of those of the vertebrate IF proteins. Furthermore, our data suggest how IF proteins might have evolved from a lamin-like ancestor in eukaryotic evolution.

Results

Isolation of a Xenopus lamin LIII gene

The genomic library used in this study was constructed by insertion of *X. laevis* genomic DNA, partially digested with restriction enzymes *Hae*III and *Alu*I, into Charon 24a phages. The library was screened with restriction fragments of a cDNA (cDNA D13) coding for *X. laevis* nuclear lamin LIII. The cloning of this cDNA has been described (Stick, 1988). Three of the isolated clones overlapped partially (Figure 1). They covered a total of ~ 25 kb of genomic DNA, and hybridized with both 5′ and 3′ end fragments of the cDNA D13 under stringent conditions. Coding regions of the gene were delineated by restriction mapping and hybridization with small cDNA fragments. The complete coding region was sequenced together with the small introns. Exon boundaries of the introns were defined by comparison with cDNA sequences. The coding region of the lamin LIII gene is split into 12 exons. The last code for two alternative C-termini of the lamin LIII proteins (see below). The 93 nucleotide long 5′ as well as the long 3′ untranslated regions are not interrupted by introns (Figure 1).

We used RNase protection assays in combination with primer extension analysis to map the transcription start site (not shown). We found a transcription start site 25 nucleotides downstream of a TATA box (underlined in Figure 2). Figure 3 shows a blot of oocyte poly(A)$^+$ RNA probed with a restriction fragment of the 3′ untranslated region (probe B in Figure 1). A RNA band of 5000 nucleotides, corresponding to the calculated transcript length (of 4721 nucleotides) plus a poly(A) tail, was detected (Figure 3a). In addition two bands of 3000 and 2100 nucleotides respectively as well as a minor band of ~ 6300 nucleotides hybridized with this probe. The shorter mRNAs are probably generated by polyadenylation at more 5′ primed polyadenylation sites (see Figure 2). The longest RNA is transcribed from a promotor upstream of the one mapped in Figure 2 since in RNase protection experiments we found no transcript exceeding the position mapped as the most downstream 3′ end of the RNA (position 4721 in Figure 2) and with all constructs used in our RNase protection experiments to map the 5′ end we found bands corresponding to the full length protected by antisense RNAs. This indicates that a mRNA exists which is transcribed from a more 5′ located promotor than that shown in Figure 2. The exact location of this promotor has not been mapped so far.

Sequence comparison between the cDNA D13 and the genomic sequence revealed three exchanges in the coding region (two silent and one missense mutations) as well as several differences in the 3′ non-coding region. The latter include 32 single base changes or single base deletions and one deletion of 23 nucleotides in the genomic sequence as well as one deletion of 44 nucleotides in the cDNA D13.

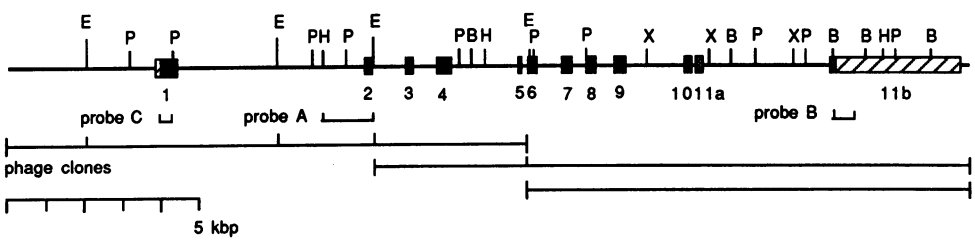

Fig. 1. Restriction map of the *X. laevis* lamin LIII gene. A restriction map of the complete lamin LIII gene is shown in the upper line: the filled boxes representing protein coding regions, the dotted boxes representing the alternative exons 11a and 11b and the hatched boxes representing the 5′ and 3′ untranslated exonic regions. The positions of the probes used in genomic Southern analysis (probes A and C) and Northern analysis (probe B) are indicated as horizontal bars. The lower lines represent the inserts of the overlapping Charon 24a phages carrying the lamin LIII sequences. Only the *Eco*RI restriction sites are indicated in the phage maps. Abbreviations of the restriction enzyme names are as follows: B, *Bam*HI; E, *Eco*RI; H, *Hin*dIII; P, *Pst*I; X, *Xba*I.

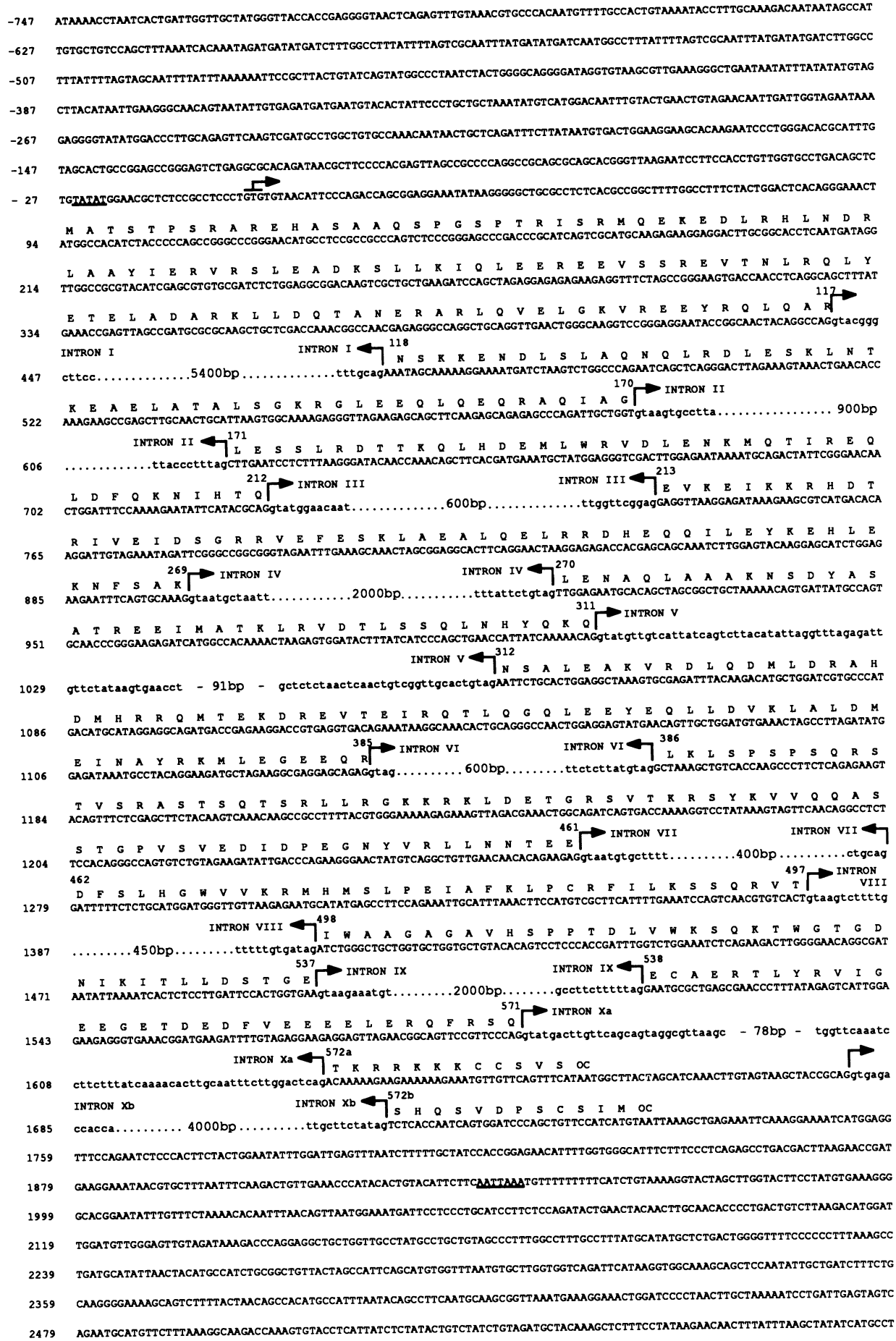

Fig. 2. Nucleotide sequence of the lamin LIII gene. The sequence of the exonic regions of the *X. laevis* lamin LIII gene and the complete sequence of the short introns V and Xa is shown as well as sequences flanking the gene at the 5′ and 3′ ends. Intronic sequences are typed in lower case letters. The size of the larger introns was defined by agarose gel electrophoresis and is depicted between dotted lines. The putative promotor sequence TATAT and polyadenylation signals are underlined. The 3′ end of the long transcripts found in Northern analysis (see Figure 3a) was mapped by RNase protection assay at position 4721. 17 nucleotides downstream of the last polyadenylation signal. It is marked by a star. Numbers to the left refer to exonic sequences taking the transcription start site which is marked by an angled arrow as +1.

By RNase protection assay we tested whether the missense mutation as well as the 44 nucleotide deletion in the cDNA D13 were present in cellular RNA from oocytes. Neither of the mutations could be detected. Since the cDNA library, the genomic library and the cellular RNA all were obtained from animals of different breeding stocks, these differences can readily be explained by the naturally occurring polymorphism and/or by errors introduced during the *in vitro* cDNA synthesis. A high degree of polymorphism in different populations of *Xenopus* has indeed been documented (Jeffreys *et al.*, 1980).

Two different C-termini of lamin LIII are generated by alternative splicing

In addition to the cDNA D13 encoding lamin LIII we have isolated another cDNA (F4-2) which encodes an isoform of LIII, differing only in the last 12 amino acids. Where comparable the two cDNAs were found to be identical with the exception of eight nucleotide exchanges probably reflecting naturally occurring polymorphism. Between codons 571 and 572 cDNA F4-2 contains an additional 77 nucleotides which are not found in cDNA D13. This insertion codes for 12 amino acids followed by a stop codon and a further 41 3′ untranslated nucleotides. The 77 extra nucleotides are encoded by a small exon located in intron 10 (see Figures 1 and 2). The generation of the two mRNA versions from the LIII gene can be explained by alternative splicing. A mRNA corresponding to cDNA F4-2 is generated when the introns flanking exon 11a are spliced out separately, while a mRNA corresponding to cDNA D13 is created by splicing out the entire intron 10 including the 77 nucleotide long exonic region. To decide whether cDNA F4-2 represents a mature mRNA that normally occurs in cellular RNA, we constructed an antisense RNA probe containing a portion of the coding region common to both cDNAs adjacent to the 77 nucleotides specific for cDNA F4-2 (see Figure 3c). Total cellular RNA was probed with this internally labelled antisense RNA in a RNase protection assay. A major protected band of 248 nucleotides (Figure 3b: band c in lanes 3 and 4) corresponding to the D13-like mRNA and a minor fragment of 360 nucleotides (Figure 3b: band b in lanes 3 and 4) were detected. The latter corresponds to a mRNA represented by cDNA F4-2. This result shows that both types of mRNA are present in oocytes. However, D13 mRNA is far more abundant than F4-2 mRNA. Nuclear lamins show the *ras*-related sequence motif CaaX at their C-terminus. This sequence is involved in post-translational isoprenylation of the cysteine residue. The cDNA F4-2 codes for the C-terminal sequence CSVS which resembles the previously defined consensus sequence CaaX.

The lamin LIII gene is present in two copies in the X. laevis genome

Xenopus laevis seems to be a tetraploid species with respect to DNA content (Bisbee *et al.*, 1977). It was therefore of interest to determine whether the number of lamin LIII genes present in the *X. laevis* genome reflects the genome duplication. A genomic blot was hybridized with a probe prepared

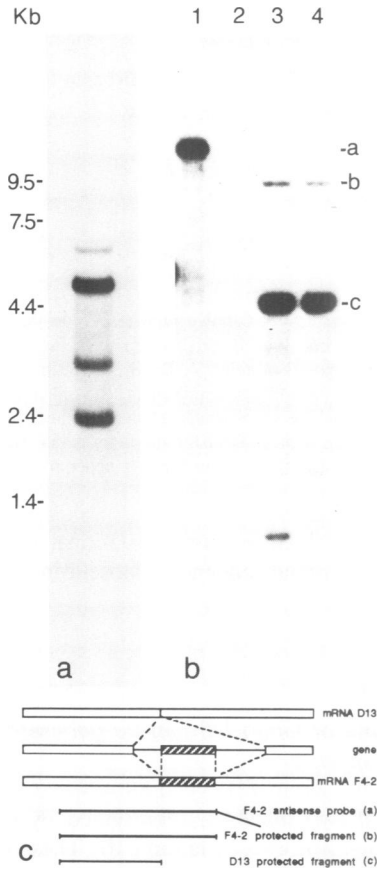

Fig. 3. Analysis of lamin LIII transcripts. **(a)** The size of lamin LIII mRNAs were estimated by Northern blot analysis. 5 μg of poly(A)$^+$ oocyte RNA were separated on a denaturing gel transferred to nylon membrane and hybridized with probe B as depicted in Figure 1. The length of LIII mRNAs were calculated to ~ 6300, 5000, 3000 and 2100 nucleotides by comparison with a RNA ladder (BRL) which was run in a parallel lane. The mRNAs of 3000 and 2100 nucleotides in length represent mRNAs with shorter 3′ untranslated regions as shown by hybridization of identical blots with more 5′ located cDNA fragments, while the mRNA of 6300 nucleotides possesses a longer 5′ untranslated region (see text). **(b)** Two differently spliced mRNAs are present in oocytes. An antisense probe (F4-2 antisense probe) containing 267 nucleotides of the coding region common to both cDNAs adjacent to the 77 nucleotides specific for cDNA F4-2 was hybridized with either 5 μg of total oocyte RNA (lanes 3 and 4) or 5 μg of yeast tRNA (lanes 1 and 2). Samples 2−4 were digested with a mixture of RNase A and RNase T1 and the protected fragments were separated on a 5% acrylamide/urea gel. (**a**) represents the undigested probe which contains additional 38 nucleotides of vector sequences, (**b**) and (**c**) represent protected fragments corresponding to cDNA F4-2 and D13 respectively as outlined in the schematic drawing. The fast migrating band in lane 3 is not found when RNase digestion is carried out at 23°C (lane 4) instead of 37°C (lane 3).

from the genomic clone as shown in Figure 1 (probe A). To avoid cross-reaction with related lamin genes we chose a probe containing mainly intronic sequences. Genomic DNA was digested with restriction enzymes for which sizes of genomic fragments could be predicted from the restriction map of the isolated clones. In each of these four cases hybridization lights up two bands of comparable intensity (Figure 4, lanes 1−4). The fragments representing the cloned gene are indicated by arrowheads in Figure 4. In contrast, when genomic DNA was hybridized with a probe containing exclusively coding sequences (probe C in Figure 2) four bands light up. Two of these bands showed a slightly higher signal intensity (Figure 4, lane 6). We have recently

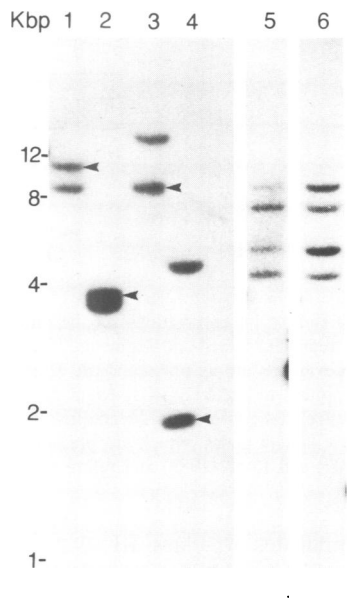

Fig. 4. Two copies of the lamin LIII gene are present in the *X. laevis* genome. *X. laevis* genomic DNA (7.5 μg/lane) was restricted with *Hin*dIII (lanes 1, 5 and 6), *Eco*RI (lane 2), *Bam*HI (lane 3) and *Pst*I (lane 4). Southern blots of the restricted DNA were hybridized with cloned, ^{32}P-labelled genomic fragments. Lanes 1−4 show hybridization with probe A (see Figure 1). Arrows point to the hybridizing DNA fragments representing the cloned gene from which the probe was prepared. Lane 6 shows hybridization with probe C (see Figure 1), and lane 5 hybridization with the corresponding probe prepared from the lamin LIII related gene (see text). Note the slight difference in relative intensity between the two pairs of bands in lanes 5 and 6, which allows the assignment to the two pairs of genes. A 1 kb ladder (BRL) was run as size marker in a parallel lane.

isolated partial genomic clones and cDNA clones coding for a LIII-related lamin. This gene shows regions of 90% sequence identity with lamin LIII in the coding region as well as in the 5′ and 3′ untranslated regions resulting in cross-hybridization of these two genes even under high stringency conditions. When a corresponding probe of the lamin LIII-related gene is used for hybridization the same four bands hybridized although in this case the relative intensity of the hybridization signals of the two pairs is reversed (Figure 4, lane 5). This allows the assignment of the two pairs of bands to the two types of genes. The observed hybridization pattern indicates that each type of lamin gene exists in at least two copies per haploid genome. This result is in agreement with the assumption that a genome duplication occurred in an ancestor of *X. laevis*.

The gene structure of lamin LIII and cytoplasmic IF genes are similar

The gene structure of all four classes of vertebrate IF proteins has previously been analysed. With the exception of neurofilament genes, which diverge totally, IF genes show a remarkable conservation of intron positions in the central rod domain (Osborn and Weber, 1986; Steinert and Roop, 1988). The conservation is even more pronounced when members of the same class are compared (Quax *et al.*, 1983, 1985; Balcarek and Cowan, 1985). In contrast, no obvious conservation of the gene structure is found in the tail domains between members of different vertebrate IF types. Despite characteristic differences, nuclear lamins share sequence principles common to all IF proteins and show moderate

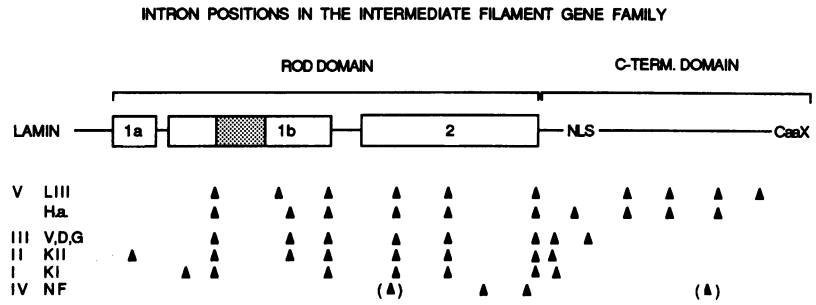

Fig. 5. A comparison of intron positions for the different types of IF genes. Intron positions (triangles) are shown with respect to the lamin protein structure. Boxes represent coiled-coil segments. The seven heptad long segment present only in lamins and invertebrate IF proteins is shown as dotted box. Abbreviations are as follows: LIII, *X. laevis* lamin LIII; H.a., *H. aspersa* non-neuronal IF protein; V, D, G, vimentin, desmin, glial fibrillary acidic protein respectively; KII, KI, neutral–basic and acidic keratins respectively; NF, neurofilament proteins; NLS, nuclear localization signal; CaaX, *ras*-related consensus sequence involved in post-translational modification of lamins. The arrowheads set in parenthesis indicate intron positions in neurofilament genes, which are present only in one of the three analysed neurofilament genes. Data for the types I–IV IF genes were taken from Steinert and Roop (1988) and references therein. Data for the *H. aspersa* IF gene are from the accompanying paper by Dodemont *et al.* (1990)

sequence similarities to class III as well as to class I vertebrate IF proteins (McKeon *et al.*, 1986; Fisher *et al.*, 1986; Weber *et al.*, 1988, 1989). Figure 5 shows a comparison of the gene structure of lamin LIII with those of IF proteins. There is a clear similarity of intron positions in the rod domain between vertebrate types I, II and III IF proteins and the lamin gene. Strikingly, five of six introns of the central coil domain are positioned at precisely the same site with respect to the codon position in lamin LIII and class III IF proteins. Intron 1 adjoins a sequence of six heptads which is not found in vertebrate IF proteins. Intron 2, which is conserved in the other IF genes, is positioned 10 codons further N-terminal in the lamin LIII gene. The tail domain of the lamin LIII gene is interrupted by four introns. None of the positions matches an intron position in the vertebrate IF genes.

Nuclear lamin LIII and the mollusc IF gene show striking similarity in their rod as well as in their tail domains

The analysis of a gene coding for a non-neuronal cytoplasmic IF protein of *Helix aspersa* reported by Dodemont *et al.*, (1990) in the accompanying paper, allows a comparison between the lamin LIII gene and an invertebrate IF protein. This comparison is even more informative. Intron positions in the central coil domain of the *H. aspersa* gene are essentially the same as in vertebrate class III IF genes. Of the four introns found in the *H. aspersa* IF tail domain three are located in regions where nuclear lamin and the invertebrate IF protein sequences can be aligned on the base of corresponding amino acid motifs (Weber *et al.*, 1989). All three introns exactly match those in the lamin LIII gene. This is strong evidence in favour of an earlier speculation, that these regions are homologous between lamins and invertebrate IF proteins. The two intron positions in the tail region, which do not match between the *Xenopus* lamin and the *H. aspersa* IF gene, are located in regions where no homologous sequence exists in the corresponding gene. Intron 7 in the *H. aspersa* gene is located in the region where the nuclear localization signal and the sites of mitotic phosphorylation are found in lamins. These signal sequences are not found in cytoplasmic IF proteins. The last exon of the lamin gene limited by intron 10 has no counterpart in the *H. aspersa* gene. It encodes the C-terminus with the CaaX motif involved in post-translational modification of lamins. Such a signal sequence is absent in cytoplasmic IF proteins.

Discussion

Two isoforms of lamin LIII can be generated by alternative splicing

Analysis of the lamin LIII gene allows one to explain the generation of two mRNAs, represented by two cDNAs, coding for two isoforms of lamin LIII. These two mRNAs can be generated by alternative splicing of the primary transcript of the LIII gene. The two cDNAs were previously isolated by expression cloning in our laboratory. They code for two lamin LIII polypeptides differing only in their 12 C-terminal amino acids. The presence of both of these mRNAs in oocytes has been proven by RNase protection analysis. Whether both are translated *in vivo* is not yet clear. For such an analysis antibodies would be necessary, which specifically recognize the extreme C-terminus of these isoforms. It is noteworthy however, that both mRNAs encode lamin isoforms carrying the CaaX motif (CSIM in cDNA D13 and CSVS in cDNA F4-2). We therefore speculate that both forms are functional *in vivo*. Alternative splicing has been suggested for the generation of human lamins A and C from the same gene. However, in this case the corresponding gene has not yet been characterized. In contrast to the situation described above human lamin C lacks the CaaX motif, and injection experiments with *in vitro* synthesized protein have shown that human lamin C on its own cannot associate with the nuclear envelope (Krohne *et al.*, 1989).

Two genes encoding lamin LIII exist in the X. laevis genome

Comparative studies of karyotypes and DNA content indicate that the various *Xenopus* species constitute a polyploid series in the proportion 2:4:8:12 (Thiébaud and Fischberg, 1977; Tymowska and Fischberg, 1982). *X. laevis* is tetraploid with respect to DNA content but functionally diploid as evidenced by the exclusive occurrence of bivalent chromosomes in meiosis (Müller, 1974). *X. laevis* is therefore referred to as 'tetraploid derived' (Graf, 1989). A duplication of the whole genome some 30 million years ago has been proposed

(Bisbee *et al.*, 1977). This would imply that most of the *X. laevis* genes should be present in two copies in the haploid genome and this holds in fact for the α- and β-globin gene cluster (Jeffreys *et al.*, 1989; Hosbach *et al.*, 1983) and for three different α-actin genes (Stutz and Spohr, 1986). The data presented here are consistent with this view.

Nuclear lamin and IF genes have conserved intron positions

Establishment of the evolutionary relationship of the IF proteins cannot be based solely on amino acid sequence comparison. Protein design constraints like the heptad sequence principle present also in other filament forming proteins result in a high degree of sequence identity with proteins, which are not considered to be members of the IF protein family. The interaction between B-type lamins and IF proteins observed in vertebrates (Georgatos and Blobel, 1987) might have led to convergent evolution of these proteins. Therefore, a rigorous proof of the evolutionary relationship of IF proteins needs additional and independent evidence as for example similarities in their gene structure. The gene structure is not under a direct adaptive selection and therefore might reflect the evolutionary history of a protein family more directly. The most striking finding of our work is the marked similarity between the structure of the genes encoding *Xenopus* nuclear lamin LIII and *H. aspersa* IF protein A. These findings do not only establish the previously suggested classification of lamins as nuclear IF proteins but furthermore allow detailed conclusions about the evolution of the IF protein family, which could not have been deduced from protein sequence comparison alone.

In regions where the amino acid sequences of the proteins are comparable, intron positions correlate precisely between the lamin gene and the gene of *H. aspersa* IF protein A with only one exception, namely intron 2. The conservation of intron positions has been maintained despite the divergence in amino acid sequence. The highest identity values are found in the short coil 1a of the rod domain, which shows ~ 40% amino acid identity, while in other regions identity values never exceed 20% and thus could not be considered as significant. These findings demonstrate that the comparably low sequence similarities between lamins and invertebrate IF proteins are indeed due to common ancestry rather than to convergent evolution as has been previously suggested.

A comparison between the gene structure of lamin LIII and that of vertebrate IF proteins shows that the lamin gene is closest to type III IF proteins. In contrast to the invertebrate IF protein discussed above, the homology between nuclear lamins and vertebrate IF proteins is restricted to the central rod domain. In the tail domain of these genes no correspondence in intron positions is found. Therefore, we believe that sequence similarities of short segments in the tail region between lamins and vertebrate IF proteins noticed previously (Fisher *et al.*, 1986) might not be based on common ancestry of these domains. Furthermore, the gene structure of the tail domains is not conserved between different types of vertebrate IF proteins. It is therefore tempting to speculate that the different classes of IF proteins evolved by combination of different non-homologous tail domains by exon shuffling and that this event happened late probably during vertebrate evolution [for more details see Discussion of the accompanying paper by Dodemont *et al.* (1990)].

Lamins are probably the ancestral members of the IF protein family

IF proteins are differentially expressed during development and cell differentiation, and are probably involved in special functions related to the differentiated state of cells. The lack of IF proteins in certain cells (Venetianer *et al.*, 1983; Giese and Traub, 1985; Hedberg and Chen, 1986; Bartnik and Weber, 1989) demonstrates that they do not serve essential housekeeping functions. Furthermore, it is not clear whether all eukaryotes possess intermediate filaments. In contrast, a nuclear lamina is ubiquitous in eukaryotes. A lamina has been demonstrated in a wide variety of organisms diverse as protozoa (Pappas, 1956), slime molds (Stick and Schwarz, unpublished), plants (Cerezuela and Moreno Diaz de la Espina, 1990), molluscs (Maul *et al.*, 1984), insects (Smith and Fisher, 1984) and vertebrates. Therefore the nuclear lamina seems to be a universal feature of eukaryotes and might have emerged early in eukaryotic evolution during the transition from the prokaryotic state. Furthermore, in those cases where information about the protein composition of the lamina is available, it has been shown that lamins (of the B-type) are constitutively expressed (Benavente *et al.*, 1985; Stick and Hausen, 1985; Lehner *et al.*, 1987; Steward and Burke, 1987) supporting the assumption that they serve housekeeping functions. The absence of lamins from cells in meiotic pachytene can be explained by functional constraints (Stick and Schwarz, 1983). Therefore, it is reasonable to assume that nuclear lamins represent the ancestral members of the IF protein family.

Evolution of the IF protein family

If indeed lamins are the ancestral members of the IF family, comparison of the lamin gene structure presented here with that of an invertebrate IF protein gene reported in the accompanying paper by Dodemont *et al.* (1990) immediately suggests how the family of IF proteins might have evolved from a lamin-like ancestor: deletion of two signal sequences found in lamins but not in IF proteins would convert a nuclear lamin into a cytoplasmic IF protein. These two signals are the nuclear localization signal responsible for the entry of lamins into the nucleus and the C-terminal CaaX motif (Holtz *et al.*, 1989; Krohne *et al.*, 1989). The later sequence is involved in membrane association of B-type lamins (Gerace and Blobel, 1989; Stick *et al.*, 1988) by post-translational isoprenylation at the C-terminal cysteine residue (Vorburger *et al.*, 1989a; Holtz *et al.*, 1989).

Deletion of the CaaX motif could be the result of an introduction of a nonsense mutation in (lamin) exon 10. This exon corresponds to the last exon in the *H. aspersa* IF gene. Deletion of the nuclear localization signal, on the other hand, could have been achieved by the creation of splice recognition sequences flanking the nuclear localization signal. This would then result in the removal of this signal during RNA maturation and, consequently, in the generation of a new intron. In fact, sequence alignment of lamins and invertebrate IF proteins necessitates the introduction of a gap in the IF sequence, which coincides with the position of the nuclear localization signal in lamins (Weber *et al.*, 1989). Furthermore, the invertebrate IF gene has an intron in this region which has no counterpart in the lamin gene (Figure 5). Thus while the amino acid sequence demands the deletion of the nuclear localization signal, the comparison of the gene structure points the way how this might have happened.

Mutant lamins missing the nuclear localization signal as well as the CaaX motif have been constructed by *in vitro* mutagenesis for other than evolutionary considerations. These artificial lamin−IF proteins form indeed 'tubular structures' in the cytoplasm (Holtz *et al.*, 1989).

Evolution of the vertebrate IF proteins from their invertebrate ancestors would need the deletion of six heptads in coil 1b of the rod domain. It is interesting to note that these six heptads common to lamins and invertebrate IF proteins start exactly at the point of an intron position common to most IF genes (Weber, 1986), including the lamin LIII gene. Thus, the loss of these 42 amino acids might be explained by the creation of a new splice acceptor site in exon 2, 126 nucleotides downstream of the one present in the lamin gene and the invertebrate IF gene. Only the position of intron 2 in the central domain differs between the lamin gene and the other IF genes. Analysis of further lamin genes will show whether the distinct position of this intron is a common feature of the lamin gene family or whether it is restricted to the *Xenopus* LIII gene.

Neurofilament genes show a completely different gene structure. Two alternative models have been outlined to explain this difference. The first proposed that the progenitor of the neurofilament genes lost its introns and gained new ones after divergence (Lewis and Cowan, 1986). An alternative view assumes that the ancestral IF gene had no introns and that introns were inserted separately into types I−III and type IV progenitors after divergence (Steinert and Roop, 1988). Our findings show that the gene structure of the ancestral IF gene must have resembled closely that of the lamin gene as well as type III IF genes and that this structure has been remarkably conserved over long distances of eukaryotic evolution. Therefore, the gene structure of vertebrate neurofilament genes has to be explained by loss and gain of introns according to the former hypothesis.

Materials and methods

A genomic library, constructed from partially *HaeIII/AluI* digested *X. laevis* genomic DNA inserted into Charon 24a phage (Stutz and Spohr, 1986), was a generous gift of G.Spohr (University of Geneva). Screening was done with restriction fragments of a cDNA encoding *Xenopus* lamin LIII (cDNA D13; Stick, 1988). Restriction fragments were labelled with digoxigenin-11-UTP using a DIG DNA labelling and detection kit (Boehringer, Mannheim, FRG). Hybridization was done in 5× SSC at 65°C overnight. Filters were washed three times 15 min each in 0.1× SSC, 0.1% SDS at 65°C. Detection of hybridized probes was done following the instructions of the manufacturer.

For phage purification, restriction enzyme digests, gel electrophoresis, Southern blotting, and subcloning into M13 derivatives or vectors (bluescribe M13, bluescript, Stratagene, La Jolla, USA) we followed standard procedures (Maniatis *et al.*, 1982). Sequencing was done by the dideoxy chain termination method (Sanger *et al.*, 1977).

For genomic blots DNA digested with appropriate restriction enzymes was separated on 0.7% agarose gels using a field reversal system with 0.5 s forward and 0.2 s reverse field. Restriction fragments for genomic blots were radiolabelled with [^{32}P]dCTP (3000 Ci/mmol, Amersham, Braunschweig, FRG) to high specific activity by the random primed method (Feinberg and Vogelstein, 1983) using a multiprime kit (Amersham, Braunschweig, FRG). Filters were hybridized in 2.5× SSPE at 65°C overnight and washed twice in 0.3× SSPE, 0.1% SDS and once in 0.5 M sodium phosphate pH 7.5.

Genomic DNA was isolated from blood cells. Blood was collected into 1% heparin in 0.8% NaCl and cells were washed once in 0.8% NaCl. They were than digested in 100× the packed cell volume of 10 mM Tris−HCl pH 7.5, 0,1 M EDTA, 0.5% SDS and 0.5 mg/ml proteinase K at 68°C for 2 h. After several extractions with phenol and phenol−chloroform DNA was purified on two successive CsCl gradients.

RNA techniques

Total RNA was extracted from *X. laevis* oocytes stages I−III (staged according to Dumont, 1972) as described by Krieg and Melton (1984). Poly(A)$^+$ RNA was isolated by selection on oligo(dT)−cellulose (Boehringer, Mannheim, FRG) using standard procedures (Maniatis *et al.*, 1982).

Northern blots of RNA separated on denaturating formaldehyde agarose gels was done essentially as described by Khandjian (1986) using cDNA fragments radiolabelled by the random primed method (Feinberg and Vogelstein, 1983).

RNase protection experiments were carried out according to Krieg and Melton (1987). Radiolabelled antisense RNA was synthesized from genomic DNA fragments cloned into bluescribe vector, using [^{32}P]UTP at a specific activity of 160 Ci/mmol. Labelled RNA was gel purified on 5% acrylamide/urea denaturing gels and eluted as described by Krieg and Melton (1987). Hybridization of total RNA (5 μg) or yeast tRNA (5 μg) and antisense RNA ($1-2 \times 10^4$ d.p.m.) was in 80% formamide, 0.4 M NaCl, 40 mM PIPES pH 6.4, and 1 mM EDTA at 58°C overnight. RNase digestion was carried out with 40 μg/ml RNase A and 2 μg/ml RNase T1 at 37° or 23°C for 1 h.

For primer extension experiments synthetic oligonucleotides (20 mers) were end-labelled using polynucleotide kinase 3′-phosphatase-free (Boehringer, Mannheim, FRG) and [γ-^{32}P]ATP (3000 Ci/mmol, Amersham, Braunschweig, FRG, gel purified on 16% acrylamide/urea denaturing gels and eluted into distilled water. Primer extension was carried out as described by Calzone *et al.* (1987) with two modifications. 5−10 μg poly(A)$^+$ RNA were hybridized with $1-2 \times 10^4$ d.p.m. end-labelled primer at 58°C overnight. Reverse transcription (with AMV reverse transcriptase, Amersham) was done at 50−60°C for 1 h in the absence of actinomycin D.

Acknowledgements

We would like to thank Prof. H. Jäckle (Munich) for generous support and for his encouragement and Prof. K. Weber (Göttingen) for stimulating discussion. We thank Heike Raab for her help in cDNA cloning. This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, grant Sti 98/3-1). R.S. is a Heisenberg Stipendiate of the DFG (Sti 98/1-1).

References

Aebi,U., Cohn,J.B., Buhle,L. and Gerace,L. (1986) *Nature*, **323**, 560−564.
Balcarek,J.M. and Cowan,N.J. (1985) *Nucleic Acid Res.*, **13**, 5527−5543.
Bartnik,E. and Weber,K. (1989) *Eur. J. Cell Biol.*, **50**, 17−33.
Benavente,R., Krohne,G. and Franke,W.W. (1985) *Cell*, **41**, 177−190.
Bisbee,C.A., Baker,M.A., Wilson,A.C., Hadji-Azimi and Fischberg,M. (1977) *Science*, **195**, 785−787.
Calzone,F.J., Britten,R.J. and Davidson,E.H. (1987) *Methods Enzymol.*, **152**, 611−632.
Cerezuela,M.A. and Moreno Díaz de la Espina,S. (1990) In Harris,J.R. (ed.), *Nuclear Structure and Function*. Plenum Press, New York, in press.
Dessev,G. and Goldman,R. (1990) *Int. J. Dev. Biol.*, **34**, 267−274.
Dodemont,H., Riemer,D. and Weber,K. (1990) *EMBO J.*, **9**, 4083−4094.
Dumont,J.N. (1972) *J. Morphol.*, **136**, 153−179.
Fawcett,D.W. (1966) *Am. J. Anat.*, **119**, 129−146.
Feinberg,A.P. and Vogelstein,B. (1983) *Anal. Biochem.*, **123**, 13−16.
Fisher,D.Z., Chaudhary,N. and Blobel,G. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 6450−6454.
Georgatos,S.G. and Blobel,G. (1987) *J. Cell Biol.*, **105**, 117−125.
Georgatos,S.G., Maroulakou,I. and Blobel,G. (1989) *J. Cell Biol.*, **108**, 2069−2082.
Gerace,L. and Blobel,G. (1980) *Cell*, **19**, 277−287.
Gerace,L. and Burke,L. (1988) *Annu. Rev. Cell Biol.*, **4**, 335−374.
Giese,G. and Traub,P. (1985) *Eur. J. Cell Biol.*, **39** (Suppl. 12), 12.
Graf,J.-D. (1989) *Genetics*, **123**, 389−398.
Gruenbaum,Y., Landesman,Y., Drees,B., Bare,J.W., Saumweber,H., Paddy,M.R., Sedat,J.W., Smith,D.E., Benton,B.M. and Fisher,P.A. (1988) *J. Cell Biol.*, **106**, 585−596.
Hancock,J.F., Magee,A.I., Childs,J.E. and Marshall,C.J. (1989) *Cell*, **57**, 1167−1177.
Hedberg,K.K. and Chen,L.B. (1986) *Exp. Cell Res.*, **163**, 509−517.
Holtz,D., Tanaka,R.A., Hartwig,J. and McKeon,F. (1989) *Cell*, **59**, 969−977.
Hosbach,H.A., Wyler,T. and Weber,R. (1983) *Cell*, **32**, 45−53.
Jeffreys,A.J., Wilson,V., Wood,D. and Simons,J.P. (1980) *Cell*, **21**, 555−564.
Khandjian,E.W. (1986) *Mol. Biol. Rep.*, **11**, 107−115.
Krieg,P.A. and Melton,D.A. (1984) *Nucleic Acids Res.*, **12**, 7057−7070.

Krieg,P.A. and Melton,D.A. (1987) *Methods Enzymol.* **155**, 397–415.
Krohne,G. and Benavente,R. (1986) *Exp. Cell Res.*, **162**, 1–10.
Krohne,G., Wolin,S.L., McKeon,F., Franke,W.W. and Kirschner,M.W. (1987) *EMBO J.*, **6**, 3801–3808.
Krohne,G., Waizenegger,I. and Höger,T.H. (1989) *J. Cell Biol.*, **109**, 2003–2013.
Lehner,C.F., Kurer,V., Eppenberger,H.M. and Nigg,E.A. (1986) *J. Biol. Chem.*, **261**, 13293–13301.
Lehner,C.F., Stick,R., Eppenberger,H.M. and Nigg,E.A. (1987) *J. Cell Biol.*, **105**, 577–587.
Lewis,S.A. and Cowan,N.J. (1986) *Mol. Cell. Biol.*, **6**, 1529–1534.
Loewinger,L. and McKeon,F. (1988) *EMBO J.*, **7**, 2301–2308.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Maul,G.G., Baglia,F.A., Newmeyer,D.D. and Ohlson-Wilhelm,B.M. (1984) *J. Cell Sci.*, **67**, 69–85.
McKeon,F.D., Kirschner,M.W. and Caput,D. (1986) *Nature*, **319**, 463–468.
Müller,W.P. (1974) *Chromosoma*, **47**, 283–296.
Osborn,M. and Weber,K. (1986) *Trends Biochem. Sci.*, **11**, 469–472.
Pappas,G.D. (1956) *J. Biophys. Biochem. Cytol.*, **2**, 431–435.
Peter,M., Kitten,G.T., Lehner,C.F., Vorburger,K., Bailer,S.M., Maridor,G. and Nigg,E.A. (1989) *J. Mol. Biol.*, **208**, 393–404.
Quax,W., Egberts,W.V., Hendriks,W., Quax-Jeuken,Y. and Bloemendal,H. (1983) *Cell*, **35**, 215–223.
Quax,W., van den Broek,L., Egberts,W.V., Ramaekers,F. and Bloemendal,H. (1985) *Cell*, **43**, 327–338.
Röber,R.-A., Weber,K. and Osborn,M. (1989) *Development*, **105**, 365–378.
Sanger,F., Nicklen,S., and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA.*, **74**, 5463–5467.
Smith,D.E. and Fisher,P.A. (1984) *J. Cell Biol.*, **99**, 20–28.
Steinert,P.M. and Roop,D.R. (1988) *Annu. Rev. Biochem.*, **57**, 593–625.
Steward,C. and Burke,B. (1987) *Cell*, **51**, 383–392.
Stick,R. (1988) *EMBO J.*, **7**, 3189–3197.
Stick,R. and Hausen,P. (1985) *Cell*, **41**, 191–200.
Stick,R. and Schwarz,H. (1983) *Cell*, **33**, 949–958.
Stick,R., Angres,B., Lehner,C.F. and Nigg,E.A. (1988) *J. Cell Biol.*, **107**, 397–406.
Stutz,F. and Spohr,G. (1986) *J. Mol. Biol.*, **187**, 349–361.
Thiébaud,C.H. and Fischberg,M. (1977) *Chromosoma*, **59**, 253–257.
Tymowska,J. and Fischberg,M. (1982) *Cytogenet. Cell Genet.*, **34**, 149–157.
Venetianer,A., Schiller,D.L., Magin,T. and Franke,W.W. (1983) *Nature*, **305**, 730–733.
Vorburger,K., Kitten,G.T. and Nigg,E.A. (1989a) *EMBO J.*, **8**, 4007–4013.
Vorburger,K., Lehner,C.F., Kitten,G., Eppenberger,H.M. and Nigg,E.A. (1989b) *J. Mol. Biol.*, **208**, 405–415.
Weber,K. (1986) *Nature*, **320**, 402.
Weber,K., Plessmann,U., Dodemont,H. and Kossmagk-Stefan,K. (1988) *EMBO J.*, **7**, 2995–3001.
Weber,K., Plessmann,U. and Ulrich,W. (1989) *EMBO J.*, **8**, 3221–3227.

Received on July 16, 1990; revised on August 22, 1990