The EMBO Journal vol.9 no.12 pp.4083-4094, 1990

Structure of an invertebrate gene encoding cytoplasmic intermediate filament (IF) proteins: implications for the origin and the diversification of IF proteins

Huub Dodemont, Dieter Riemer and Klaus Weber

Max Planck Institute for Biophysical Chemistry, Department of Biochemistry, D-3400 Göttingen, FRG

Communicated by K.Weber

The structure of the single gene encoding the cytoplasmic intermediate filament (IF) proteins in non-neuronal cells of the gastropod *Helix aspersa* is described. Genomic and cDNA sequences show that the gene is composed of 10 introns and 11 exons, spanning >60 kb of DNA. Alternative RNA processing accounts for two mRNA families which encode two IF proteins differing only in their C-terminal sequence. The intron/exon organization of the *Helix* rod domain is identical to that of the vertebrate type III IF genes in spite of low overall protein sequence homology and the presence of an additional 42 residues in coil 1b of the invertebrate sequence. Intron position homology extends to the entire coding sequence comprising both the rod and tail domains when the invertebrate IF gene is compared with the nuclear lamin LIII gene of *Xenopus laevis* presented in the accompanying report of Döring and Stick. In contrast the intron patterns of the tail domains of the invertebrate IF and the lamin genes differ from those of the vertebrate type III genes. The combined data are in line with an evolutionary descent of cytoplasmic IF proteins from a nuclear lamin-like progenitor and suggest a mechanism for this derivation. The unique position of intron 7 in the *Helix* IF gene indicates that the archetype IF gene arose by the elimination of the nuclear localization sequence due to the recruitment of a novel splice site. The presumptive structural organization of the archetype IF gene allows predictions with respect to the later diversification of metazoan IF genes. Whereas models proposing a direct derivation of neurofilament genes seem unlikely, the earlier speculation of an mRNA transposition mechanism is compatible with current results.

Key words: alternative RNA processing/evolution/gene structure/intermediate filament proteins/invertebrates/lamins

Introduction

In vertebrates the complex multigene family encoding the structural proteins of the cytoplasmic intermediate filaments (IF) comprises many distinct members, which show cell- and tissue-specific expression patterns. By the criteria of protein sequences and intron positions of the corresponding genes a convenient subdivision into four classes can be made (reviewed by Osborn and Weber, 1986; Steinert and Roop, 1988). All type III genes—vimentin, desmin, glial fibrillary acidic protein (GFAP) and peripherin—have essentially identical intron patterns. The epithelial keratin type I and II genes show exon/intron structures clearly related to type III genes but also possess type-specific introns (Marchuk *et al.*, 1984; Johnson *et al.*, 1985; Quax *et al.*, 1985; reviewed by Steinert and Roop, 1988). In addition certain variations within each keratin type have been found (reviewed in Bader *et al.*, 1986; Krauss and Franke, 1990). The finding that type IV neurofilament genes display a totally different structural organization gave rise to considerable speculation on the divergence of neuronal versus non-neuronal IF genes during metazoan evolution (Lewis and Cowan, 1986; Julien *et al.*, 1987, 1988; Myers *et al.*, 1987; Lees *et al.*, 1988; Steinert and Roop, 1988). Based on their cDNA sequences and ultrastructure (Aebi *et al.*, 1986; Fisher *et al.*, 1986; McKeon *et al.*, 1986; Gruenbaum *et al.*, 1988), the nuclear lamins seem to constitute together with the cytoplasmic IF proteins a superfamily, which in mammals comprises ~40 different members.

In an attempt to understand the basis of the astounding complexity of vertebrate IF proteins we started a survey of IF among invertebrates. Immunological and biochemical data on gastropods, annelids and nematodes point to a much lower IF complexity and define only two distinct IF prototypes—a neuronal and a non-neuronal (nn) type (Bartnik *et al.*, 1985, 1987a,b). The latter seems present in all nn cells known to display IF by electron microscopic criteria. IF isolated from several epithelia, as well as from the glial cells of the snail *Helix pomatia*, contain two immunologically related polypeptides of mol. wts 66 kd (A) and 52 kd (B) (Bartnik *et al.*, 1985, 1987b). A and B polypeptides purified from oesophagus epithelium form homopolymeric IF *in vitro*, while the keratin filaments of vertebrate epithelia are obligatory heteropolymers. Protein sequences documented the tripartate organization typical for vertebrate IF proteins —an α-helical rod domain with coiled-coil forming ability flanked by non-helical head and tail domains. They also showed that A and B differ only by a C-terminal extension unique to the longer A chain (Weber *et al.*, 1988). The nn IF proteins of the gastropod *H.pomatia* and the nematode *Ascaris lumbricoides* have two structural features in common with the nuclear lamins, which are absent from vertebrate IF proteins: an increase of the rod domain by 42 residues located in coil 1b and a moderate lamin homology segment of 120 residues in the tail domain (Weber *et al.*, 1988, 1989). These features provided the first direct support for earlier speculations on a common ancestry of nuclear lamins and cytoplasmic IF proteins (Osborn and Weber, 1986; Bartnik *et al.*, 1987a; Myers *et al.*, 1987; Steinert and Roop, 1988).

A direct relation between the invertebrate IF proteins and one or other of the vertebrate nn IF proteins remained unclear due to the low overall sequence identities except for the consensus sequences at the ends of the rod domain (Weber *et al.*, 1988, 1989). Since subtypes of vertebrate IF genes differ distinctly in exon/intron patterns, we have now analysed the organization of the gene(s) encoding the nn IF

proteins of *Helix aspersa*. Here we show that the two closely related IF proteins A and B are generated by alternative RNA processing pathways from the single copy nn IF gene. We compare the structure of the invertebrate IF gene with that of the different vertebrate IF genes and with the *Xenopus* gene encoding nuclear lamin LIII characterized by Döring and Stick in the accompanying report (Döring and Stick, 1990). Our results show a striking conservation of intron positions among the invertebrate nn IF and the vertebrate type III genes for the rod domains. More importantly this conservation of gene structure extends only in the lamin and the invertebrate IF genes into the tail domain. From the differences in this region we propose that the archetype cytoplasmic IF gene arose from a nuclear lamin-like ancestor by the loss of two signal sequences: the nuclear localization sequence and the CaaX motif. We further draw several conclusions on the subsequent diversification of metazoan IF genes.

Results

Characterization of non-neuronal IF mRNAs and their cDNA clones

Poly(A)$^+$ polysomal RNA from oesophagus epithelium of *H.aspersa* was fractionated by size and translated *in vitro*. The resulting products were analysed before and after immunoprecipitation with rabbit antiserum to *Helix* nn IF proteins. Proteins A (66 kd) and B (52 kd) represent authentic newly synthesized polypeptides encoded by distinct mRNAs, each of which occurs in two size classes (~4.5 and 2 kb; see Figure 1). In each class the A-encoding mRNA has a larger size than the mRNA directing the synthesis of B. The clear separation of the translational activities for A and B strongly argues against a post-translational derivation of the shorter B chain from the longer A polypeptide.

Comparison of the immunoprecipitation and *in vitro* translation profiles shows that the highest relative enrichment for IF-specific sequences resides in the 4.5 kb size fraction. This mRNA was used to screen an initial cDNA plasmid library comprising 3000 independent clones generated from total unfractionated mRNA. Fractions 6 and 7 (Figure 1A) were used separately to synthesize cDNA probes for differential hybridization of A- and B-specific sequences. Thirty-two cDNA clones, which showed medium to strong hybridization, were isolated for subsequent testing by hybrid selected translation (Figure 2A). Plasmids E_3 and E_4 each selected both mRNAs simultaneously although with different efficiency. Whereas hybrid selection with plasmid E_3 (lane 2) was heavily biased towards the B-encoding mRNA, both mRNAs were retained by plasmid E_4 (lane 3) to an extent comparable with their relative abundance in the original mRNA population (lane 4). Final evidence that the E_3 and E_4 plasmid represent the B and A mRNA sequence was obtained by sequencing the cDNA inserts. These are shown schematically in Figure 2B. Plasmids pSonnIF52 E_3 and pSonnIF66 E_4 contained the C-terminal portion of the coil 2 domain plus the complete tail domains followed by long 3′-untranslated sequences. To retrieve the rest of the protein coding region a primer extension cDNA library was established using fragment E_3B to prime the original mRNA. Clone pSonnIF PE-1 contained the entire sequence plus a 267 bp 5′-untranslated region preceding the ATG initiation codon (Figure 2B).

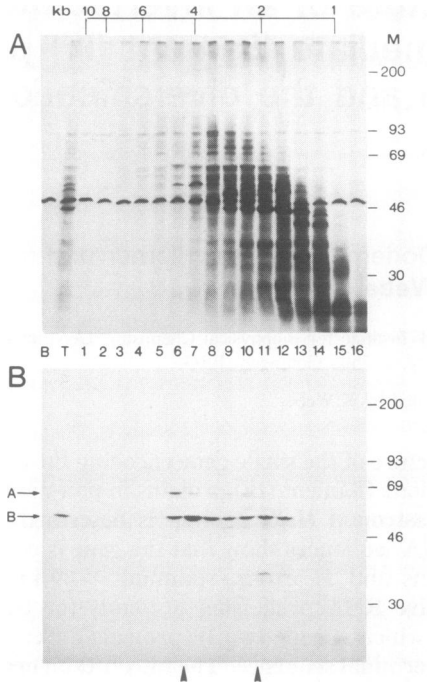


Fig. 1. Size determination of nn IF mRNA. **(A)** Poly(A)$^+$ polysomal RNA (10 μg) from oesophagus epithelium was electrophoresed in a denaturing gel. The 0.8–10 kb size range of the gel (see scale at the top) was divided into 16 fractions. Recovered mRNA was translated *in vitro*. Newly synthesized polypeptides were subjected to SDS–PAGE (lanes 1–16) and fluorography (18 h). Lane B: products endogenous to the reticulocyte lysate. Lane T: translation products of total mRNA. M: position of protein size standards (kd). **(B)** Fluorograph (4 weeks exposure) of gel electrophoretic separation of the *in vitro* translation products shown in (A) after immunoprecipitation with rabbit antibodies against *Helix* nn IF proteins. Arrowheads mark the two peaks of translational activity for A and B proteins.

The three cDNA inserts were used to screen additional cDNA libraries representing total poly(A)$^+$ polysomal RNA from oesophagus epithelium and cerebral ganglion. Each library yielded several large cDNA clones (Figure 2B), all of which contained 5′-untranslated sequences 243–284 nucleotides in length. The single open reading frames of the B and A mRNAs encode 453 and 576 amino acid residues flanked by UAG and UAA termination signals and large 3′-non-coding regions comprising 2042 and 2032 nucleotides, respectively. Both 3′-stretches contain the consensus AATAAA polyadenylation signal (Proudfoot and Brownlee, 1976; Birnstiel *et al.*, 1985) located 15 nucleotides upstream of the poly(A) addition site. Remnants of the poly(A) tail were found in several independent A-specific cDNAs but were absent from clones derived from the B-encoding mRNA which all had the same penultimate nucleotide at their 3′-ends. Both mRNAs are identical over the 5′-untranslated sequences and the open reading frames up to the codon for Ser452. At this point the A mRNA diverges from B mRNA with the codon for Thr453 instead of Ser453, which is the C terminus of the B protein. The A mRNA continues with the tail sequence unique to the A polypeptide, followed by a 3′-untranslated region totally distinct from its counterpart in B mRNA. The coexistence of two mRNAs which share long identical 5′-sequences but differ completely in their 3′-ends strongly suggests that they arise from a single gene via alternate RNA processing pathways. In addition no sequence differences were found

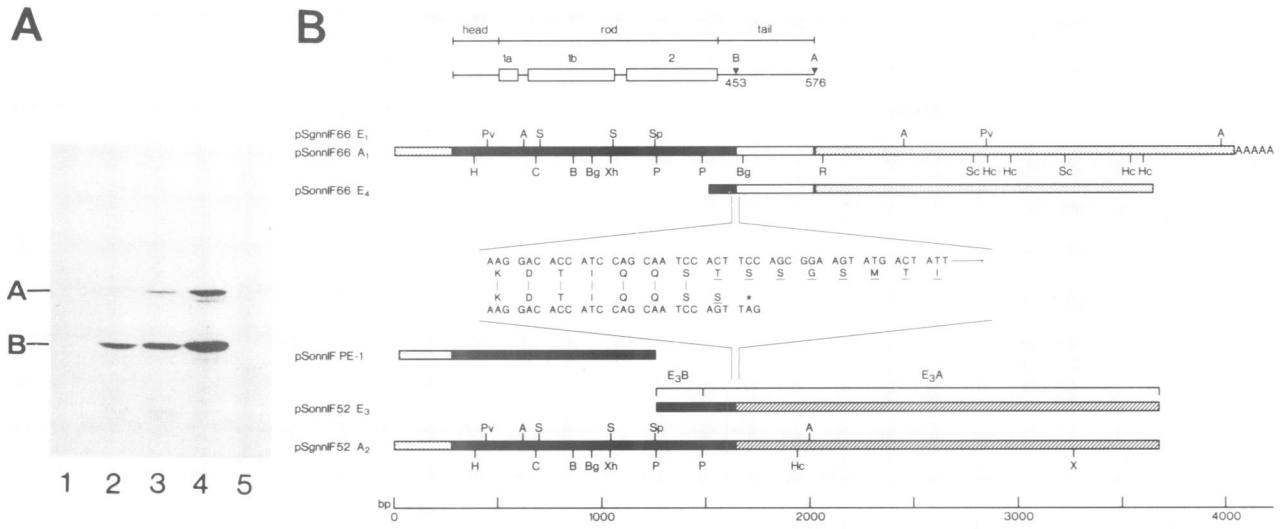


Fig. 2. Characterization of nn IF cDNAs. (**A**) Identification of cDNAs by hybrid selected translation. The fluorograph (8 weeks exposure) shows the immunoprecipitated translation products synthesized *in vitro* by total mRNA (lanes 4 and 5) and mRNAs specifically hybridized to plasmids pBR322 (lane 1), E_3 (lane 2) and E_4 (lane 3). Immunoprecipitation was with rabbit antibodies to *Helix* nn IF proteins (lanes 1–4) or the corresponding preimmune serum (lane 5). (**B**) Schematic representation of cDNAs. The three domain structure of IF proteins is presented at the top. Arrowheads in the tail region mark the C termini of protein B (453 residues) and A (576 residues). The cDNAs are grouped as A- and B-specific and are designated at the left; pSo and pSg refer to cDNAs derived from oesophagus and ganglion, respectively. Solid boxes represent protein coding sequences common to both A and B. Open boxes define the sequence for the extended tail region unique to protein A. Stippled boxes indicate 5′-untranslated sequences. Single- and cross-hatched boxes mark the 3′-non-coding regions of B- and A-specific cDNAs, respectively. Nucleotide and amino acid sequences are shown only for the regions where A and B cDNAs diverge. Restriction enzyme maps are given for both cDNA types. Sites are marked for *Acc*I (A), *Bam*HI (B), *Bgl*II (Bg), *Cla*I (C), *Eco*RV (R), *Hin*cII (Hc), *Hin*dIII (H), *Pst*I (P), *Pvu*II (Pv), *Sac*I (S), *Sca*I (Sc), *Sph*I (Sp), *Xba*I (X) and *Xho*I (Xh). A size scale (bp) is indicated at the bottom. E_3B and E_3A represent subcloned regions of the pSonnIF52 E_3 cDNA.

when the A- and B-encoding mRNAs from oesophagus epithelium were compared with their counterparts from the glia cell rich ganglion. The presence of a single gene giving rise to the same A and B proteins in both tissues is consistent with previous immunological data (Bartnik *et al.*, 1987b).

The cDNA insert from pSonnIF PE-1 was hybridized to a RNA blot of poly(A)$^+$ and poly(A)$^-$ polysomal RNA from oesophagus and ganglion. Since this probe contains predominantly protein coding sequences, all nn IF mRNAs should be detected. Figure 3 displays in poly(A)$^+$ RNA (lanes 1 and 3) two prominent transcripts of ~4.6 and 4.2 kb. By hybridization with specific probes containing exclusively 3′-untranslated sequences, the upper and lower bands could be assigned to the A- and B-encoding mRNAs, respectively (data not shown). In addition to the large mRNAs, a faint broad band is detected in the 2 kb size class of poly(A)$^+$ RNA. Apparently, the small mRNAs are much less abundant than anticipated from the cell free translation studies, indicative of their higher translational efficiency *in vitro* as compared with the large mRNAs. Surprisingly, the small RNAs are the only species detected in poly(A)$^-$ RNA (lanes 2 and 4) where they occur at considerably higher levels than in poly(A)$^+$ RNA. To enable the analysis of RNA sequences solely derived from the low mol. wt poly(A)$^+$ and poly(A)$^-$ species, new cDNA libraries were constructed from gel purified 2–2.5 kb fractions. Sequence analysis of several independent cDNA clones revealed that the small poly(A)-containing and poly(A)-deficient RNAs represent 3′-truncated variants of their large counterparts. They are contiguous with the large mRNAs throughout the entire protein coding sequences and the 5′-portions of the 3′-untranslated regions. For one representative each from the small A- and B-specific cDNAs the outermost 3′-end was determined; they were found downstream from the stop codon at 58 and 226 nucleotides, respectively.

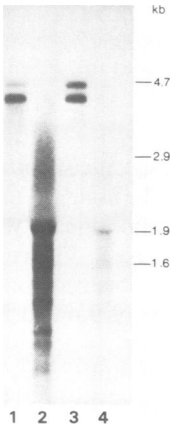


Fig. 3. Detection of nn IF mRNA by blot hybridization. Two μg of poly(A)$^+$ and 20 μg of poly(A)$^-$ polysomal RNA from oesophagus epithelium (lanes 1 and 2) and cerebral ganglion (lanes 3 and 4) were electrophoresed under denaturing conditions, transferred to nitrocellulose, hybridized to nick-translated cDNA pSonnIF PE-1 and autoradiographed (9 h exposure). Positions of RNA size standards (rat 18S and 28S rRNA, *E.coli* 16S and 23S rRNA) are indicated at the right. Due to the high concentration of rRNA in this region, the 2 kb RNA bands detected in poly(A)$^-$ RNA are distorted and shifted to lower mol. wt.

Isolation of the single non-neuronal IF gene

Hybridization studies using entire cDNA inserts (Figure 2B) as probes on Southern transfers of *H.aspersa* genomic DNA digested with various restriction enzymes yielded complex

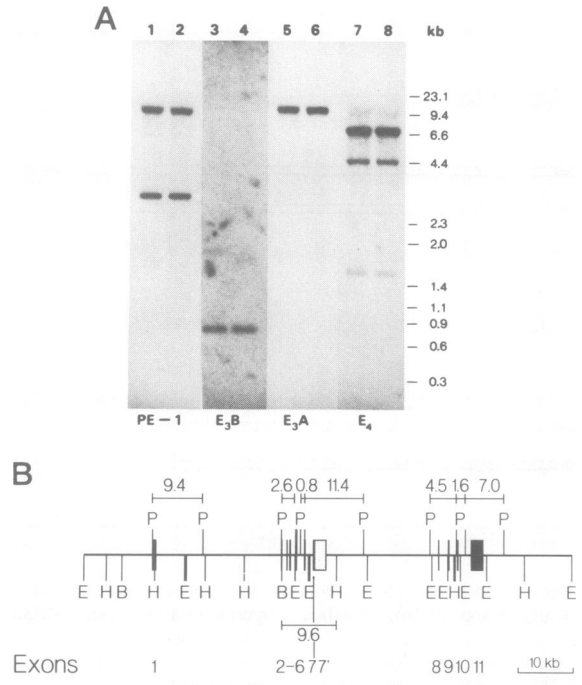


Fig. 4. Genomic representation of the nn IF gene. (**A**) Southern blot analysis of *Pst*I digested *H.aspersa* genomic DNA (5 μg per lane), hybridized to nick-translated cDNAs pSonnIF PE-1 (lanes 1 and 2), E_3B (lanes 3 and 4), E_3A (lanes 5 and 6) and pSonnIF66 E_4 (lanes 7 and 8). Autoradiography was 3 days for lanes, 1, 2, 5 and 6 and 1 week for lanes 3, 4, 7 and 8. Positions of DNA size standards (*Hin*dIII digest of phage λ DNA and *Hae*III digest of φX174 DNA) are indicated at the right. (**B**) Restriction enzyme map of the nn IF gene and position of exons. Exons are represented as solid boxes. The positions of the genomic DNA *Pst*I fragments detected in (A) and the 9.6 kb *Bam*HI−*Hin*dIII fragment (see text) are shown. Sites are marked for *Bam*HI (B), *Eco*RI (E), *Hin*dIII (H) and *Pst*I (P). A size scale of 10 kb is indicated.

band patterns. When specific subprobes were used, all bands that hybridized could be assigned to a single large gene. No additional hybridization signals were detected under the stringent conditions used. Figure 4A shows the simple hybridization pattern obtained for *Pst*I digested genomic DNA. Two fragments of 9.4 and 2.6 kb were detected by the pSonnIF PE-1 insert, whereas the subcloned E_3A and E_3B portions from pSonnIF52 E_3 hybridized to 11.4 and 0.8 kb fragments, respectively. The entire pSonnIF66 E_4 insert revealed three fragments of 7.0, 4.5 and 1.6 kb, but also, though weakly, the 11.4 kb fragment due to a small 131 bp sequence overlap with the E_3A probe. All seven *Pst*I fragments were cloned separately. Their position within the overall genomic DNA organization was established using appropriate subprobes in comparative Southern blot hybridization analyses of single and multiple digests with *Bam*HI, *Eco*RI, *Hin*dIII and *Pst*I. An eighth 0.16 kb *Pst*I fragment, which escaped detection by the E_3B probe, was isolated on a 9.6 kb *Bam*HI−*Hin*dIII fragment bridging the gap between the 2.6 and 0.8 kb *Pst*I fragments. In the resulting map of 87 kb (Figure 4B) the snail nn IF gene comprises at least 60 kb.

Sequence and organization of the non-neuronal IF gene

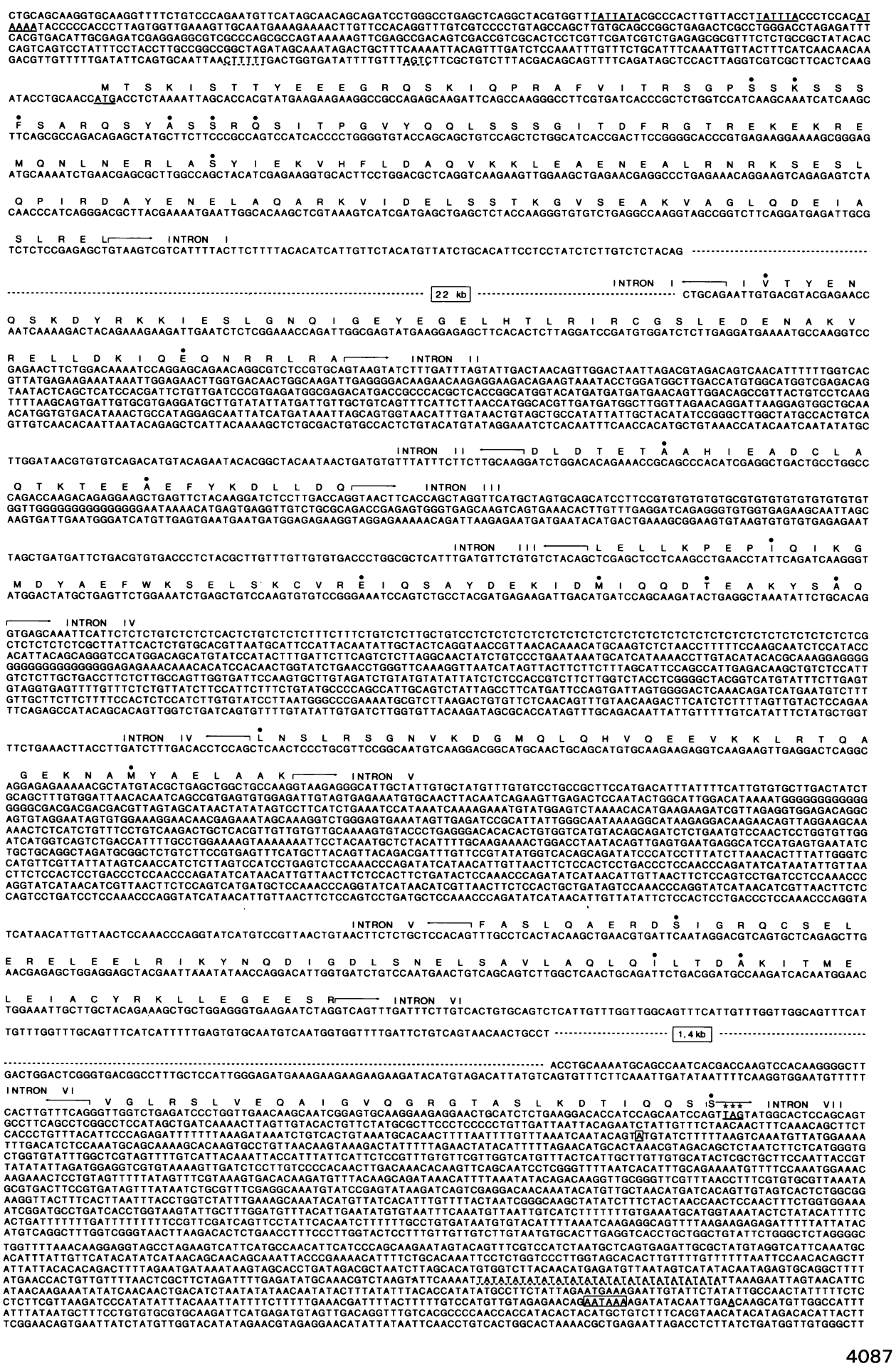
Alignment of detailed restriction enzyme maps of the cloned genomic DNA fragments and cDNAs, combined with Southern hybridization analysis showed that the entire cDNA

Table I. Sequences at the exon/intron boundaries of the nn IF gene

	5′		3′
Consensus	NNNAG gtaagt	.....yyyyyyyyyyncag	GNNNN
Intron		Kb	
	161		*162*
	GluLe		uIle
I *o	GAGCT gtaagt...	(~ 22)	...ctgcag AATTG
	223		*224*
	Ala		Asp
II ○	GTGCA gtaagt...	(0.745)	...tgcaag GATCT
	255		*256*
	Gln		Leu
III *o	ACCAG gtaact...	(0.393)	...ctacag CTCGA
	308		*309*
	Gln		Leu
IV *o	CACAG gtgagc...	(0.994)	...ctccag CTCAA
	350		*351*
	Lys		Phe
V *o	CCAAG gtaaga...	(1.345)	...ccacag TTTGC
	424		*425*
	SerAr		gVal
VI *o	TCTAG gtcagt...	(~ 1.4)	...tttcag GGTTG
	453		
	SerS er***		
VII	ATCCA GTtagt...		
	453		*454*
	SerT		hrSer
VII	ATCCA gttagt...	(~ 22)	...ttgcag CTTCC
	492		*493*
	Lys		Thr
VIII *	CCAAG gtatga...	(~ 1.6)	...ttacag ACCCA
	527		*528*
	Thr		Ile
IX *	ACACC gtaagt...	(~ 1.5)	...tcgcag ATCTG
	565		*566*
	Asn		Glu
X *	GAAAC gtaagc...	(~ 2.3)	...ttgcag GAGAA

Sequences are aligned with consensus donor (left) and acceptor (right) splice sites (Padgett *et al.*, 1986). Exon sequences are given in capital letters. Sequences pertaining to intron 7 are presented twice to emphasize the alternative splicing pathways (see Figure 6A). Amino acids are given by the three-letter code to highlight the codon phase used by the introns. Numbers refer to residue positions in the nn IF polypeptides. Precise or estimated lengths (kb) of the introns are given in parentheses. Asterisks and circles refer to identically positioned introns for the *Xenopus* lamin LIII gene (Döring and Stick, 1990) and the hamster vimentin gene (Quax *et al.*, 1983), respectively.

sequences were contained within the series of eight *Pst*I fragments. The gene is composed of 11 exons and 10 introns (Figure 4B), which interrupt the protein coding sequence at the positions summarized in Table I. All exons and four of the introns (numbers 2−5) were fully sequenced. Only short stretches of the other introns adjacent to the exons were determined, except for intron 7, of which 3.7 kb of the 5′-sequence was analysed. Figure 5 shows the combined sequence data (~14 000 nucleotides) including 1 kb of 3′-flanking sequence. The introns vary widely in size, ranging from 393 bp (intron 3) to 22 kb (introns 1 and 7). All exon/intron boundaries (Table I) are consistent with the consensus 5′ and 3′-splice sites (Padgett *et al.*, 1986). Sequence comparison of the exons with the previously determined cDNAs revealed only four minor discrepancies. These are restricted to untranslated regions and can be attributed either to polymorphisms or artefacts introduced during reverse transcription.



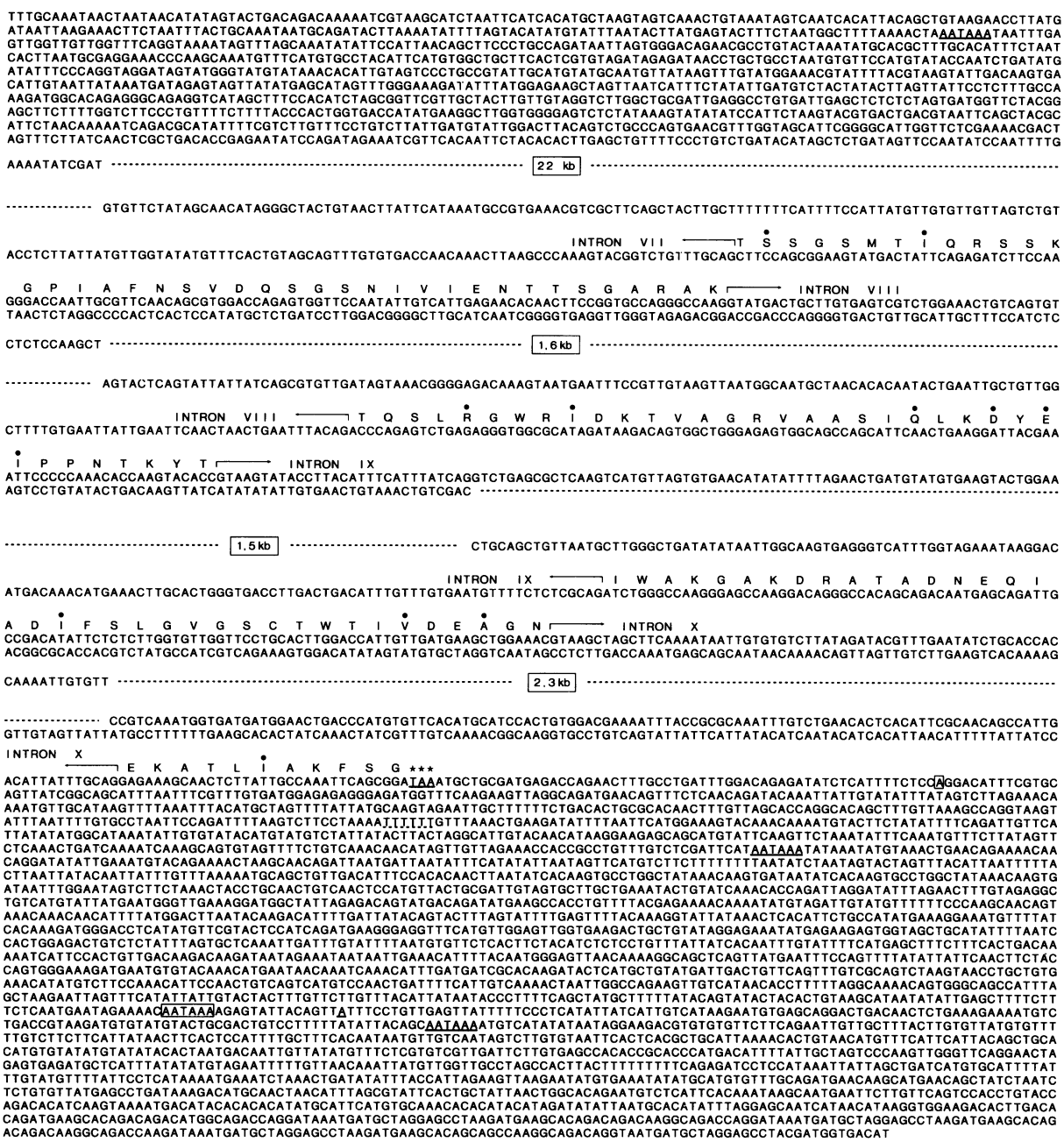


Fig. 5. Nucleotide and predicted amino acid sequence of the nn IF gene. Per line 120 nucleotides are given. Introns are delineated by arrows and marked by Roman numerals. Sizes are indicated only for those introns which were not fully sequenced (see also Table I). The ATG start codon is underlined. The two stop codons TAG and TAA are underlined and marked by asterisks. Amino acid residues are represented by single letters above the middle position of each codon. Dots above residues indicate substitutions with respect to the directly determined sequence of *H.pomatia* nn IF proteins (Weber *et al.*, 1988). Canonical AATAAA polyadenylation signal sequences in intron 7 (i.e. exon 7′, see text) and exon 11 are indicated. Framed hexamer sequences are utilized by the major RNA transcripts; other hexamers are underlined. Sites of poly(A) addition are found 15 nucleotides downstream from the framed signals and are marked by the underscored A nucleotide. Boxed A nucleotides are located 228 and 58 nucleotides downstream from the TAG and TAA stop codons. These designate the 3′-ends of a representative small B and A RNA transcript, respectively. A cluster of three TATA-like sequences, displayed in the first line of the sequence, is underlined (see text). Discrepancies between cDNA and genomic sequences are indicated by dashed lines. The genomic CTTTTT and AGTC stretches (upstream of the ATG start codon) were represented in cDNA as CCTTTT and AGTTC, respectively. The genomic TTTTTT hexamer sequence (~350 nt downstream from the TAA stop codon in exon 11) was found in cDNA as a pentamer. The 17-fold TA repeat sequence in genomic DNA (located in intron 7) was found in different copy numbers in several cDNAs.

The open reading frame starts with the ATG codon at nucleotides 613–615. It is located within a sequence differing only in four out of 13 positions from the consensus sequence GCCGCCACCATGG, considered to be a favourable context for translational initiation (Kozak, 1989). The predicted *H.aspersa* nn IF protein sequence is in excellent agreement with the sequence directly determined for *H.pomatia* (Weber *et al.*, 1988, 1989). Only 33 amino acid substitutions are found, accounting for <6% interspecies divergence. The sequence upstream of the ATG start codon shows a stretch of three TATA-box related elements (Breathnach and Chambon, 1981), surrounded by GC rich

sequences reminiscent of the GC boxes found in promoter regions of several higher eukaryotes (Mitchell and Tjian, 1989). To locate the presumptive transcription initiation site, a primer extension cDNA library was constructed from oesophagus poly(A)$^+$ RNA primed by a restriction fragment whose 5′-end mapped ~200 bp downstream from the TATA region. The resulting cDNA clones contained an extended sequence contiguous with the entire upstream genomic DNA sequence and reaching beyond the *Pst*I site which marks the start of the sequence in Figure 5. This shows that the true promoter region was not isolated and that the TATA-like elements are an integral part of an exceptionally long 5′-leader sequence which comprises at least 612 bp. The largest 5′-untranslated sequence obtained by the initial cDNA cloning was only 284 bp. The additionally acquired 5′-sequence taken together with a poly(A) tail length of ~100 nucleotides compensates well for the 0.4−0.5 kb size discrepancies between the largest cDNAs and the corresponding mRNAs.

Expression of the non-neuronal IF gene: alternative RNA processing pathways

The nn IF mRNAs which differ only at their 3′-ends arise from the single gene by differential utilization of polyadenylation sites (Figure 6A). The mRNA encoding the B protein is generated from the short putative primary transcript B which terminates within intron 7. Processing occurs by splicing of the first six exons to exon 7 which at its 3′-end links up to the sequence derived from the 5′-end of intron 7 (e.g. exon 7′), yielding for the B-encoding mRNA one additional serine codon after the serine codon common to both mRNAs. Splicing of exon 7 to exon 8 from the long putative precursor A abolishes the penultimate serine codon and the adjacent 3′-untranslated sequence of the B-encoding mRNA. Instead, a threonine codon is generated which marks the onset of the extended tail region unique to A. Three additional splices of exons 8−11 complete the sequence of the A-encoding mRNA. Selection of polyadenylation sites located upstream of the major sites probably accounts for the origin of the minor 2−2.5 kb RNA species.

Various tissues known to express the nn IF proteins (Bartnik *et al.*, 1985, 1987b) were tested by Northern analysis. The large A and B mRNAs occur in all 14 tissues tested, although at very different levels (Figure 6B and C). Furthermore, the proportion of the two major mRNAs varies from a large bias towards the B mRNA in the 'albumen gland' (lane 8) to an almost 1:1 ratio in foot sole epidermis (lane 7). This tissue-specific regulation of expression probably reflects different rates of synthesis and/or processing of the long and short primary transcripts. No apparent size differences can be seen for each of the large mRNAs among the tissues analysed. Therefore selection of polyadenylation signal sequences, other than those shown to be functional in oesophagus and ganglion (see boxed AATAAA hexamers in Figure 5), does not occur in the case of the major transcripts.

The structures of the invertebrate nn IF gene and a vertebrate nuclear lamin gene are highly related

Alignment of the different sequences shows a highly similar intron pattern for vertebrate type III IF genes and the invertebrate nn IF gene (Figures 7 and 8). The strict conservation of the first six intron positions (see also Table

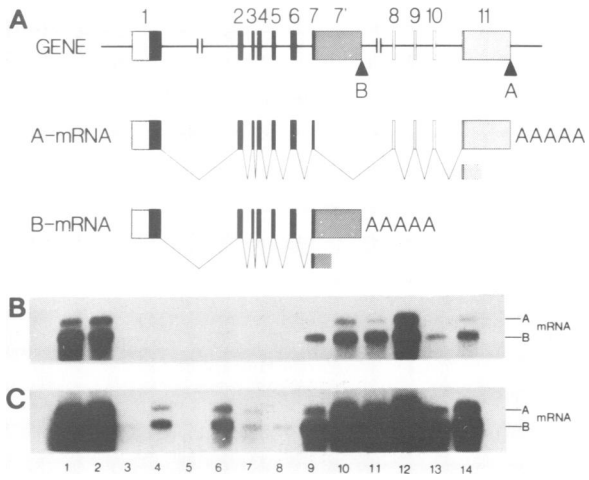


Fig. 6. Expression of the nn IF gene. (**A**) Diagrammatic representation of splicing pathways. Exons are represented as boxes and drawn to scale. Designation of coding and non-coding sequences is the same as for cDNAs (see Figure 2). Arrowheads mark polyadenylation signals for the major A and B transcripts. Stippled open boxes below exons 7′ and 11 indicate 3′-ends of minor transcripts, which use alternative polyadenylation sites, located upstream of the major sites. (**B** and **C**) Northern analysis of nn IF mRNAs in various tissues. Equal amounts (2 μg) of poly(A)$^+$ polysomal RNA from different tissues were subjected to blot hybridization with nick-translated pSonnIF PE-1. RNA was isolated from oesophagus (1), ganglion (2), lung (3), digestive gland (4), kidney (5), heart (6), foot sole epidermis (7), albumen gland (8), oviduct (9), receptaculum seminis (10), mucus gland (11), vas deferens (12), penis (13) and dart sac (14). Autoradiography was 10 h (B) and 50 h (C). Only the 4.5 kb size region is shown.

I), which interrupt the coding sequence of the rod domain, is quite remarkable considering the low protein sequence homology (~23% over the rod) and the presence of an extra 42 residues in the coil 1b of the invertebrate protein. Beyond the rod domain there is no obvious alignment of the shorter tail domains of vertebrate type III IF proteins and their two introns with the much longer tail of the *Helix* protein A and its four introns. Figure 8 shows that the intron patterns of the vertebrate keratin I and II genes differ to various extents from the common six intron pattern of type III/nn IF genes and that vertebrate neurofilament genes have an entirely different organization (Lewis and Cowan, 1986; Julien *et al.*, 1987, 1988; Myers *et al.*, 1987; Lees *et al.*, 1988).

The homology in structure is even more impressive for the invertebrate nn IF gene and the vertebrate lamin gene. Eight of the 10 introns occur at the same position in the two genes. Of the six introns interrupting the coding sequence of the rod, all but intron 2 are identically placed. Thus although lamins and invertebrate IF proteins share the same sized insert in coil 1b, which starts past intron 1, the next intron is differently placed. Intron 2 of the nn IF gene occupies the same position as in type III IF genes, while the corresponding lamin intron occurs 30 nucleotides upstream. The coding sequence of the tail domains is interrupted by four introns, with lamin introns 7−9 corresponding to introns 8−10 of the nn IF gene. The alignment of Figure 7, based on earlier sequence data of three invertebrate nn IF proteins (Weber *et al.*, 1988, 1989), shows that the unique intron 7 of the nn IF gene occupies a position which in the lamin gene corresponds to the sequence encoding the nuclear localization signal (Loewinger

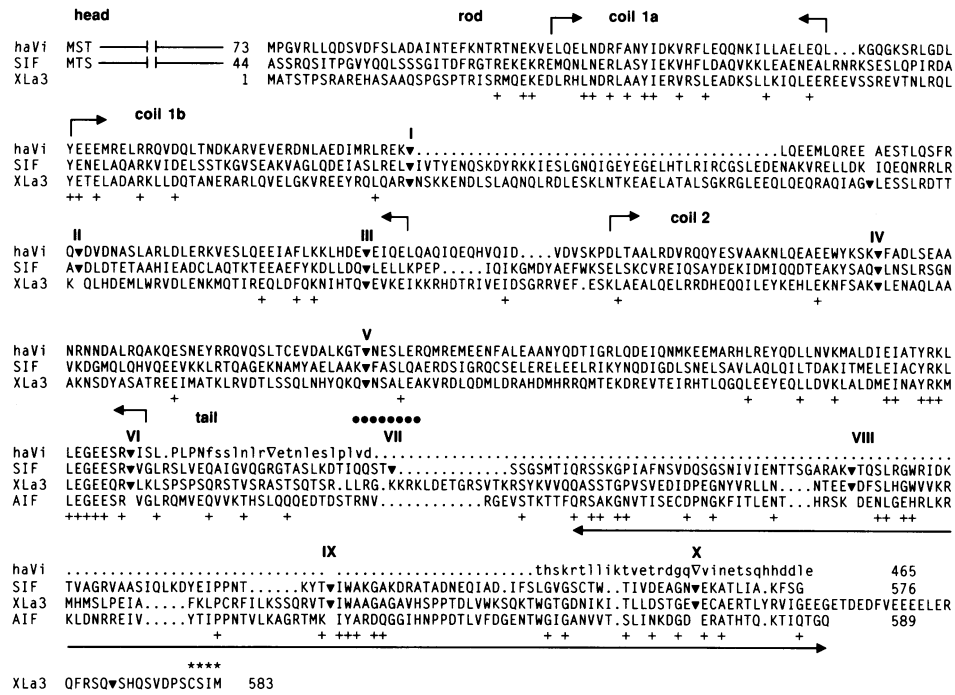


Fig. 7. Comparison of the *Helix* nn IF, hamster vimentin and *Xenopus* lamin LIII gene organizations. Predicted protein sequences of the *Helix* nn IF protein A (SIF; Figure 5), hamster vimentin (haVi; Quax *et al.*, 1983) and *Xenopus* lamin LIII (XLa3; Döring and Stick, 1990) are aligned essentially as in previous protein comparisons (Weber *et al.*, 1988, 1989) and the intron positions (arrowheads) are added. Common structural principles and the domains of IF proteins are indicated (for nomenclature see Geisler and Weber, 1982; Steinert and Roop, 1988). Except for the consensus sequences at both ends of the rod domains, sequence principles rather than actual sequences are conserved. Plus signs mark identical residues in all three rod domains and in the two SIF/XLa3 tail domains. Lower case letters used in the haVi tail domain indicate uncertain homology versus the other two genes. The horizontal arrow delineates the homology region in the tail domains of lamins and invertebrate IF proteins. Here the alignment includes the corresponding sequence of an *Ascaris* IF protein (AIF) which displays a higher similarity to lamin sequences (Weber *et al.*, 1989). The introns of the *H.aspersa* nn IF gene are marked by Roman numerals. Note that the intron patterns of the *Helix* and the hamster vimentin genes are identical over the rod domain in spite of low overall protein sequence homology and the 42 residue deletion in coil 1b of vimentin. Over the rod domain the lamin gene displays the same pattern except for the position of intron 2. This striking similarity in organization continues in the tail domains of the *Helix* nn IF and the lamin gene. Note that introns 8–10 of the nn IF gene correspond, both in position and phase (see Table I), to lamin introns 7–9. The position of the unique intron 7 of the nn IF gene corresponds to the region encoding the nuclear localization signal sequence in the lamin gene (see dots). Lamin intron 10, which has no counterpart in the nn IF gene, separates the last exon, which carries the CaaX motif (asterisks) from exon 10, whose 5′-sequence is aligned with the C-terminal end of the *Helix* protein. Sequence alignment of seven lamins including that of *Drosophila* (Pollard *et al.*, 1990) shows in the regions corresponding to exons 8 and 9 of lamin LIII a variability in length of two and one amino acids, respectively. The corresponding exons 9 and 10 of the *Helix* nn IF gene differ by one and two residues in length from their LIII counterparts. In line with the higher lamin homology of the *Ascaris* IF proteins (Weber *et al.*, 1989) the corresponding regions of this invertebrate IF protein show no length variability versus lamin LIII.

and McKeon, 1988). In fact, in earlier protein alignments a deletion had to be introduced in this region. Lamin intron 10, more than half of the preceding exon 10, as well as exon 11, have no counterparts in the nn IF gene. The high relation in structure of the two genes is remarkable in view of the relatively low protein sequence identity (22% over the rods and 21% over the homology region in the tail domains; for a higher value of 32% in the tail of the *Ascaris* IF protein see Figure 7 and Weber *et al.*, 1989).

Discussion

Characteristics of expression of the non-neuronal IF gene

We characterized the gene which encodes the nn prototype of cytoplasmic IF of the invertebrate *H.aspersa*. The gene is represented in the genome as a single large copy spanning at least 60 kb and is structurally organized into 11 exons and 10 introns, all of which interrupt the protein coding sequence. Using exon-derived fragments as probes for hybridization under stringent conditions, no signals other than those belonging to the gene could be detected. The absence of closely related sequences conforms to previous immunological data showing that several distinct antibodies defined only a single nn IF type in a variety of tissues (Bartnik *et al.*, 1987b). The nn IF prototype comprises two proteins, A (66 kd) and B (52 kd), which are contiguous from amino acid residues 1 (Met) through to 452 (Ser). Proteins A and B diverge at position 453 (Thr453 for A, Ser453 for B), which marks the C terminus of B and the onset of the extended tail domain of protein A comprising an additional 123 residues.

Proteins A and B each are encoded by multiple mRNAs constituting two distinct families, which arise from the single gene via alternative RNA processing. Selective utilization of polyadenylation sites and differential RNA splicing pathways produce mRNAs with identical 5′-sequences but divergent 3′-ends. The last four exons, which specify the extended tail domain sequences, are either retrieved or eliminated during processing, giving rise to mature A- or B-encoding mRNA, respectively. Similarly featured multiple RNAs produced from a single transcription unit are known

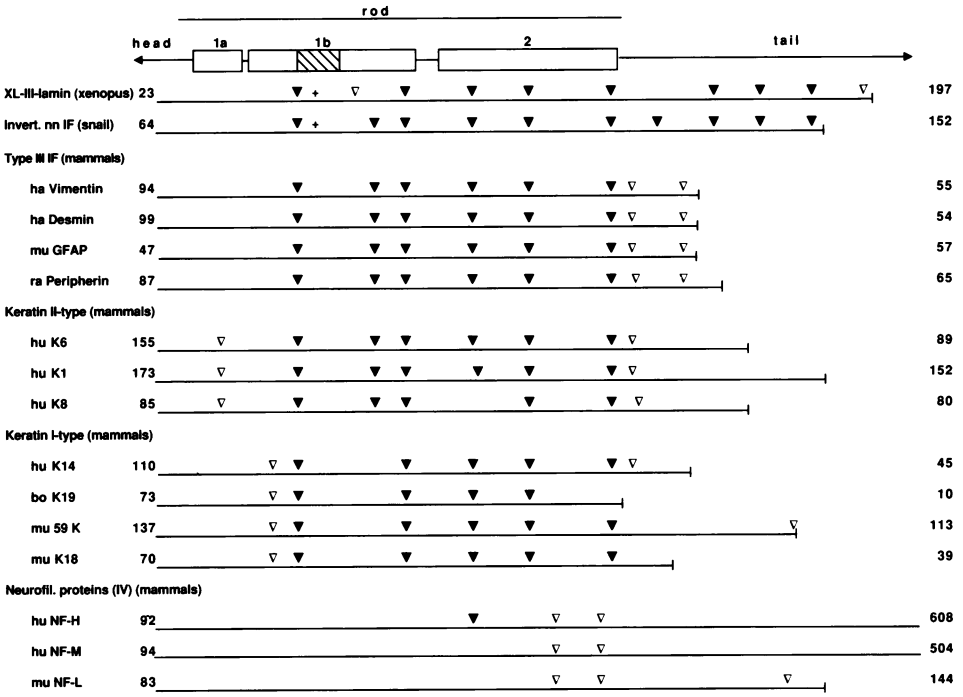


Fig. 8. Summary of intron positions in the lamin/IF multigene family. The tripartate protein structural organization is indicated at the top. Residue numbers for the variable head and tail domains are given at the sides. The hatched box in coil 1b marks the extra 42 residues (six heptads) unique to lamins and invertebrate IF proteins (plus sign). Solid arrowheads specify intron positions corresponding to those of the *Helix* nn IF gene. Non-related introns are marked by open arrowheads. Note that only the four vertebrate type III IF genes and the invertebrate nn IF gene display strict conservation of intron/exon patterns over the rod domains. Keratin type I and II genes show at least one type-specific intron (first introns) and some variation in each subfamily. For minor shifts of some keratin introns, not detectable on the scale of this graph, see the summaries of Bader *et al.* (1986) and Krauss and Franke (1990). Vertebrate neurofilament genes have an entirely different organization although the first intron of the NF-H gene is nearly in the same position as intron 4 of the vertebrate type III IF genes. The recently discovered vertebrate gene for the new IF protein nestin, typically found in neuroepithelial cells and the stem cells of the central nervous system (Lendahl *et al.*, 1990) shows the two intron positions common to all other vertebrate neurofilament genes. While this relation was not recognized in the original report it has been noted in a parallel review (Steinert and Liem, 1990). Intron positions are given for the genes encoding *Xenopus* lamin LIII (Döring and Stick, 1990); *Helix* nn IF (this work); hamster vimentin (Quax *et al.*, 1983); hamster desmin (Quax *et al.*, 1985); mouse GFAP (Balcarek and Cowan, 1985); rat peripherin (Thompson and Ziff, 1989); human keratin 6 (Tyner *et al.*, 1985); human keratin 1 (Johnson *et al.*, 1985); human keratin 8 (Krauss and Franke, 1990); human keratin 14 (Marchuk *et al.*, 1984); bovine keratin 19 (Bader *et al.*, 1986); murine keratin 59 kd (Krieg *et al.*, 1985); murine keratin 18 (Ichinose *et al.*, 1988); human NF-H (Lees *et al.*, 1988); human NF-M (Myers *et al.*, 1987); mouse NF-L (Lewis and Cowan, 1986). Note that only in the lamin and the *Helix* nn IF genes the homology in organization extends from the rod domain over the tail domain.

for several eukaryotic genes (for a review, see Leff *et al.*, 1986). In particular the gene encoding the secreted and membrane bound forms of immunoglobulin IgM heavy chain shows structural and expression characteristics (Alt *et al.*, 1980; Early *et al.*, 1980; Rogers *et al.*, 1980), very similar to those described here for the *Helix* nn IF gene. For both genes utilization of the 5′-proximal polyadenylation site results in a precursor RNA molecule terminating within an intron. As a consequence no downstream acceptor splice site is available for the 3′-located exon, which instead extends with the adjacent 5′-sequences retrieved from the intron. In contrast, selection of the farther 3′-located polyadenylation site, which in the *Helix* nn IF gene lies ~28 kb downstream from the first, leads to elimination of the entire intron from the transcript by splicing to a distally located exon.

The A- and B-encoding RNA families each comprise a single large polyadenylated mRNA, which is the major species, and a minor group of small RNA transcripts of heterogeneous size. The latter RNAs represent 3′-truncated variants of the large mRNAs and are found predominantly in the poly(A)$^-$ fraction. The small proportion with a poly(A) tail probably originates from precursors which selected alternative poly(A) sites upstream of the canonical AATAAA sequence utilized by the major transcripts (for a review see Birnstiel *et al.*, 1985). The origin of the non-polyadenylated RNAs, which constitute the large majority of the small transcripts, is not certain. The 3′-non-coding regions of the large mRNAs encoding A and B contain 64 and 48 copies, respectively, of the sequence motifs TATT or ATTT. These are believed to represent signals for cell-mediated RNA degradation (Shaw and Kamen, 1986; Brawerman, 1987; Wilson and Treisman, 1988; Hennessy *et al.*, 1989). Therefore, specific degradation at the 3′-ends of the large mRNAs may well account for the occurrence of the small poly(A)-deficient RNAs.

Evolutionary derivation of cytoplasmic IF proteins

The recent elucidation of invertebrate IF protein sequences emphasized two distinct structural features shared by the lamins but absent from all vertebrate IF proteins currently known: six additional heptads in coil 1b and a long homology sequence in the tail domains (Weber *et al.*, 1988, 1989). These characteristics provided direct molecular support for a common ancestral origin for cytoplasmic IF and lamin genes as was postulated earlier (Osborn and Weber, 1986; Bartnik *et al.*, 1987a; Myers *et al.*, 1987; Steinert and Roop,

1988). However, conclusive evidence about the exact nature of molecular relationships and consequently the presumptive evolutionary history of the cytoplasmic IF protein/nuclear lamin gene superfamily can only be retrieved from gene structure analysis. The data presented in our study and in the accompanying report of Döring and Stick (1990) reveal a remarkable similarity in the structural organization of the genes encoding the *Helix* nn IF proteins and the *Xenopus* nuclear lamin LIII. Eight out of 10 introns present in both genes are located at homologous positions and the common principles of organization cover both the rod and tail domains. The invertebrate IF/lamin relationship in gene organization is much stronger than could have been anticipated from the low overall protein sequence homology. Conversely, even among evolutionarily highly conserved proteins like the α- and β-tubulins and the actins, sequence homology is not paralleled by a strict conservation in structural organization of the corresponding genes (for a review, see Dibb and Newman, 1989).

It is generally believed that the nuclear lamina is a ubiquitous component of the eukaryotic nuclear envelope (Fawcett, 1966, 1981). Although direct molecular proof of nuclear lamins in protozoa and plants is still lacking (for ultrastructural studies see Pappas, 1956; Beams *et al.*, 1957; Mercer, 1958; Cerezuela and Moreno Dias de la Espina, 1990), the recent biochemical evidence for yeast lamins (Georgatos *et al.*, 1989) implies that nuclear lamins are a very early acquisition of eukaryotic life. The strikingly similar structural organizations of the invertebrate nn IF gene and the lamin gene not only considerably substantiates their common ancestry. The few differences in intron/exon patterns also immediately suggest how the archetype IF protein gene arose from a lamin-like progenitor. While analysis of additional lamin genes will have to show whether the shift in intron 2 is a feature common to all lamins, the two other differences in gene organization have direct functional impact. Intron 7 of the *Helix* nn IF gene has no counterpart in the lamin gene. It occurs in a region which in the lamin gene encodes the nuclear localization signal, a functional prerequisite for entry of lamins into the nucleus (Loewinger and McKeon, 1988; Holtz *et al.*, 1989). Conversely, intron 10 which delineates the last exon of the lamin gene is absent in the nn IF gene. This short exon ends with the CaaX motif which is involved in a complex post-translational cascade creating a membrane binding site necessary for functional integrity of B-type lamins (Holtz *et al.*, 1989; Vorburger *et al.*, 1989; for *ras* proteins see Hancock *et al.*, 1989). Whereas the nuclear localization signal was lost by acquisition of a new splice site, the CaaX sequence could have been removed by the introduction of a stop codon to shorten the protein chain. The elimination of these two signal sequences freed the lamin-like archetype IF protein from nuclear compartmentalization as well as unwanted membrane interactions and provided the possibility to form cytoplasmic IF. This hypothesis conforms to the structural appearance of the lamina in *Xenopus* oocytes (Aebi *et al.*, 1986). It is also in line with the ability of lamins to form IF-like filaments *in vitro* (Aebi *et al.*, 1986) and with the properties of certain lamin mutants constructed by *in vitro* mutagenesis for functional experiments (Loewinger and McKeon, 1988; Holtz *et al.*, 1989). Lamins with mutated signal sequences form 'tubular filamentous structures' in the cytoplasm. Since plants are thought to have cytoplasmic IF (Hargreaves *et al.*, 1989), the origin of the lamin/IF divergence probably occurred already in early eukaryotic evolution.

Divergence of metazoan IF proteins

In spite of the extra 42 residues in the coil 1b domain and a low level of sequence identity with vertebrate type III proteins in the rod domain, the invertebrate nn IF gene displays its first six introns precisely at the same positions as all four type III IF genes. This conservation of gene organization over the rod domain holds with the exception of intron 2 also for the lamin gene. During evolution the position and number of introns interrupting the tail domain changed in the vertebrate type III genes versus the invertebrate nn IF gene, which kept the close relation to the lamin gene. Interestingly, the gene structure of tail domains is also not conserved between different classes of vertebrate IF genes (Figure 8). This may imply that vertebrate IF protein diversity evolved by different combinations of the rod domain with distinct tail domains due to exon shuffling.

Our results limit the number of models which can account for the divergence of metazoan IF and the emergence of type I and II (keratin) and type IV (neurofilament) genes. They argue against the speculation that IF genes evolved from an intronless primordial gene with subsequent lineages acquiring distinct intron positions (for a general discussion of various models see Steinert and Roop, 1988). While a progenitor for type I–III genes was postulated from common features in exon/intron patterns (Marchuk *et al.*, 1984; Johnson *et al.*, 1985; Quax *et al.*, 1985) we now see that vertebrate type III genes are more closely related to the invertebrate nn IF gene and the lamin gene than are the keratin type I and II genes. We speculate that keratin genes arose later in metazoan evolution from the type III/nn IF lineage but evolved at a faster rate than type III genes. Of the various models which have tried to explain the completely different structure of vertebrate neurofilament genes (Lewis and Cowan, 1986; Julien *et al.*, 1987, 1988; Myers *et al.*, 1987; Lees *et al.*, 1988; Steinert and Roop, 1988) only two are in principle compatible with the organization of the archetype IF gene derived by us and by Döring and Stick (1990). The first model proposes that a mRNA transposition event abolished the ancient introns and a few type specific introns were subsequently acquired by the neurofilament gene(s) (Lewis and Cowan, 1986). The second assumes that intron 1 of the NF-H gene, which is not present in the other neurofilament genes, still marks a direct derivation from the archetype IF gene (Julien *et al.*, 1988; Lees *et al.*, 1988). Since intron 1 of the NF-H gene is shifted in position and changed in phase versus the 'corresponding' intron in the *Helix* nn IF and *Xenopus* lamin genes, we consider it a later acquisition and favour the model of Lewis and Cowan (1986) for the original derivation of a neurofilament gene. However, the structure of an invertebrate neurofilament gene will be necessary to evaluate this model directly.

Materials and methods

Animals

H.aspersa rather than *H.pomatia* was used since it is available year round from Pacific Biomarines Laboratories, Venice, CA, USA. Oesophagus, ganglion and various other tissues were dissected, frozen immediately in liquid nitrogen and stored at −80°C.

Isolation of polysomal RNA

Polysomal RNA was prepared by the method of Palmiter (1974) with a few modifications. Briefly, frozen tissue (10 g maximum) was homogenized at 0°C with a Polytron blender in a 20-fold excess of hypotonic buffer containing 50 mM Tris−HCl, pH 8.0, at 0°C, 25 mM NaCl, 5 mM $MgCl_2$ and 2% (v/v) Triton X-100. After removal of nuclei and cell debris by low speed centrifugation the supernatant was brought to a final concentration of 1 mg/ml heparin and 100 mM $MgCl_2$ and kept at 0°C for 1 h. The precipitate was collected by centrifugation at 0°C, 30 000 *g* for 30 min. Pellets were resuspended in 50 mM Tris−HCl, pH 7.5, at 37°C, 50 mM CDTA (1,2-cyclohexylenedinitrilotetraacetic acid monohydrate, Sigma), 0.5% Na-sarcosyl and 100 μg/ml proteinase K (Boehringer). Following incubation at 37°C for at least 1 h, polysomal RNA was precipitated with ethanol. Poly(A)$^+$ RNA was isolated (Aviv and Leder, 1972) via two cycles of affinity chromatography to oligo(dT)−cellulose (type 7, Pharmacia).

Identification of nn IF encoding mRNA

Poly(A)-containing polysomal RNA was electrophoresed in a 1.5% agarose gel in the presence of 10 mM methylmercury hydroxide (Bailey and Davidson, 1976). The region of the lane with RNA species ranging in size from 0.8 to 8 kb was sliced into 3 mm sections and mRNA was recovered (Dodemont *et al.*, 1985). For preparative purposes tRNA carrier was omitted. Individual RNA fractions were assayed in a nuclease-treated rabbit reticulocyte lysate in the presence of L-[^{35}S]methionine (1200 Ci/mmol, Amersham). *In vitro* translation products were analysed in SDS−10% polyacrylamide gels either directly or following immunoprecipitation with a rabbit antibody to *Helix* nn IF proteins (Weber *et al.*, 1988). Gels were fluorographed to Kodak XAR-5 film.

Sizing and polyadenylation of poly(A)-deficient RNA

RNA which did not bind to oligo(dT)−cellulose at high ionic strength was rechromatographed several times to ensure complete depletion of poly(A)-containing RNA. The resulting poly(A)-deficient RNA was size fractionated and the 2−2.5 kb size class recovered without carrier as outlined above. Four μg of this RNA was polyadenylated *in vitro* at 37°C for 1 h in a 50 μl reaction mixture containing 50 mM Tris−HCl, pH 8.0, at 37°C, 250 mM NaCl, 10 mM $MgCl_2$, 10 mM DTT, 500 μg/ml BSA, 100 μM ATP, 0.5 μCi [α-^{32}P]ATP (3000 Ci/mmol, Amersham) and 1 U of polyadenylate nucleotidyltransferase (Sippel, 1973). The reaction was stopped by addition of CDTA, Na-sarcosyl and proteinase K to final concentrations of 20 mM, 0.5% and 100 μg/ml, respectively. After 1 h at 37°C the reaction product was applied to a 2.5 ml P-60 gel filtration column (Bio-Rad) in a siliconized Pasteur pipette and eluted in 10 mM Tris−HCl, pH 8.0, 100 mM NaCl, 1 mM CDTA, 0.1% Na-sarcosyl. Eluted RNA was finally phenol extracted and ethanol precipitated. Denaturing gel electrophoretic analysis showed no obvious degradation of the newly polyadenylated RNA which had acquired an average 3′-poly(A) track of 80 residues.

cDNA synthesis

Oligo(dT) primed cDNA synthesis from either total poly(A)-containing polysomal RNA or size fractionated poly(A)$^-$ RNA, polyadenylated *in vitro*, was performed essentially as described (Gubler and Hoffman, 1983). RNA (2 μg) was reverse transcribed at 42°C for 15 min in a 40 μl reaction mixture containing 100 mM Tris−HCl, pH 8.3, at 42°C, 4 mM Na-pyrophosphate, 8 mM $MgCl_2$, 10 mM DTT, 100 μg/ml oligo(dT)$_{12-18}$, 1 mM of each dNTP, 10 μCi [α-^{32}P]dCTP (3000 Ci/mmol) and 50 U of reverse transcriptase (Amersham). The reaction was terminated and the cDNA−mRNA hybrid products recovered via gel filtration as detailed above. Second strand synthesis was carried out in a 100 μl volume consisting of 50 mM Tris−HCl, pH 7.6, at 15°C, 100 mM KCl, 10 mM $(NH_4)_2SO_4$, 10 mM $MgCl_2$, 10 mM DTT, 100 μg/ml BSA, 100 μM of all four dNTPs, 2.5 μCi [α-^{32}P]dCTP and 150 μM βNAD. *Escherichia coli* RNase H (Amersham), DNA polymerase I (Boehringer) and DNA ligase (Biolabs) were added at 10, 250 and 10 U/ml, respectively and the mixture was incubated sequentially at 12 and 18°C for 1 h each. The reaction was stopped by addition of CDTA to 5 mM and subsequent heat inactivation of the enzymes at 70°C for 15 min. The pH was adjusted to 8.3 at 37°C followed by addition of fresh $MgCl_2$, DTT, BSA and each dNTP to concentrations of 10 mM, 5 mM, 50 μg/ml and 50 μM, respectively. The cDNA was made blunt-ended with 4 U of T4 DNA polymerase (Amersham) at 37°C for 30 min. Following another heat inactivation step in the presence of 20 mM CDTA the reaction mixture was treated with RNase A at 20 μg/ml, 37°C, 30 min. Finally, Na-sarcosyl and proteinase K were added to concentrations of 0.5% and 100 μg/ml, respectively and incubation was continued for 1 h. The cDNA was isolated via gel filtration, phenol extraction and ethanol precipitation as above.

Construction and screening of cDNA libraries

The initial library was established with oligo(dC) tailed cDNA, annealed to *Pst*I cut, oligo(dG) tailed pBR322 vector (Peacock *et al.*, 1981). DNA was used to transform competent AG-1 cells (Stratagene) which were plated on SOB/tetracycline agar media (Hanahan, 1985). Only 3000 transformants were picked in microtitre plates and stored in glycerol at −80°C. In all other experiments, blunt-end cDNA was ligated in a 1:1 molar ratio to gel purified, *Sma*I cut, dephosphorylated pUC18 vector followed by transformation of JM109 (Yanisch-Perron *et al.*, 1985). A protocol based on the method of Hanahan (1985) routinely yielded 2 × 10^8 transformants/μg of supercoiled vector DNA. Colonies were grown to small size at high density on nitrocellulose filters (Schleicher and Schuell) on SOB/ampicillin agar media. Libraries constructed this way comprised 1−2 × 10^6 transformants/μg input cDNA and <2% background. Colonies were replicated and lysed as described (Hanahan and Meselson, 1980) followed by hybridization as detailed below. The initial library was probed with cDNA prepared from gel purified mRNA templates enriched for nn IF-specific sequences (see above) using random calf thymus DNA primers for reverse transcription. All other screenings were done with nick-translated cDNA inserts (Rigby *et al.*, 1977). All probes were labelled to high specific activity (0.5−1 × 10^9 c.p.m./μg) using [α-^{32}P]dCTP (3000 Ci/mmol). Positive clones were purified via one or two additional rounds of screening. Isolation and restriction enzyme analysis of plasmid DNAs followed standard procedures (Maniatis *et al.*, 1982).

Synthesis and cloning of primer extension cDNA

The 222 bp *Pst*I fragment E_3B derived from cDNA clone pSonnIF52 E_3 comprising amino acid residues 325−398 was used to prime the synthesis of cDNA on total poly(A)-containing polysomal RNA from oesophagus. Denatured DNA (100 ng) and mRNA (40 μg), corresponding to an estimated 10-fold molar excess of DNA with respect to the mRNA target sequence, were coprecipitated and resuspended in 40 μl hybridization solution containing 80% formamide, 50 mM Na-PIPES, pH 6.4, 400 mM NaCl, 5 mM CDTA and 0.2% Na-sarcosyl. The mixture was covered with paraffin oil, heated at 90°C for 10 min and incubated at 52°C for 16 h allowing the formation of RNA−DNA hybrids (Casey and Davidson, 1977). After chilling on ice, the volume was adjusted with H_2O and 5 M NaCl to 200 μl 250 mM NaCl, followed by three extractions with chloroform:isoamyl alcohol (24:1) to remove paraffin oil and recovery of nucleic acids by ethanol precipitation. Synthesis and cloning of cDNA was performed as above except that for the chimera formation the oligo(dC):(dG) joining procedure was used (Peacock *et al.*, 1981). All clones generated hybridized to the nick-translated primer fragment. Analysis of 72 randomly chosen clones yielded cDNA pSonnIF PE-1. This contained adjacent to intact E_3B a 1239 bp extension, 267 bp of which were 5′-untranslated sequences. For further extension towards the 5′-end a 413 bp *Hinc*II−*Hin*dIII fragment isolated from the cloned 5′-portion of the gene was used as a primer. This contained the last 308 nucleotides of the 5′-untranslated region next to the first 105 nucleotides of coding sequence. Experimental conditions were as above except that the library was established with blunt-end ligated cDNA. Colony hybridization was performed with a nick-translated 212 bp *Pst*I−*Hin*fI fragment which represents the outermost 5′-genomic DNA sequence.

Positive hybrid selected translation

Specific mRNAs were hybrid selected from total poly(A)$^+$ polysomal RNA by 10 μg denatured plasmid DNAs bound to nitrocellulose filters in 50 μl reaction volumes containing 65% formamide, 50 mM Na-PIPES, pH 6.4, 400 mM NaCl, 5 mM CDTA, 0.2% Na-sarcosyl, 100 μg/ml yeast tRNA and 200 μg/ml mRNA. Mixtures were heated at 70°C for 10 min followed by incubation at 52°C for 3 h. Initially filters were washed twice for 30 min each with hybridization solution without RNA. Further washings and subsequent elution and recovery of hybrid released mRNAs were as reported (Dodemont *et al.*, 1985). The mRNAs were identified by gel electrophoretic analysis of immunoprecipitated *in vitro* translation products.

RNA and DNA blot hybridization

Transfer of alkali-treated DNA and glyoxylated RNA from agarose gels to nitrocellulose membranes was as described (Southern, 1975; Thomas, 1980). Filters were prehybridized at 42°C for 4 h in 50% formamide, 50 mM Na-phosphate, pH 6.8, 1 mM Na-pyrophosphate, 5 × SSC, 5 mM CDTA, 5 × Denhardt's solution and 100 μg/ml denatured salmon sperm DNA. For hybridization this mixture was replaced by a fresh solution of the same composition except that it contained 20 mM Na-phosphate, 1 × Denhardt's and 5 ng/ml boiled nick-translated cDNA insert (see above). After 18−40 h incubation at 42°C filters were washed twice for 1 h at 42°C with hybridization solution without probe, followed by 0.5 × SSC−0.1%

SDS and 0.1 × SSC−0.1% SDS at 50−60°C for 15 min each. Filters were autoradiographed to Kodak XAR-5 film using intensifying screens.

Genomic DNA analysis

Genomic DNA was isolated from several tissues using standard techniques (Maniatis *et al.*, 1982). Despite careful preparation none of the various DNAs seemed to have a very high molecular size (>100 kb) thus precluding the option of generating cosmid libraries. Instead, several partial genomic DNA libraries were established in parallel, each enriched (50- to 200-fold) for a *Pst*I fragment hybridizing to selected cDNA subprobes (see text). Gel purified, *Pst*I cut, dephosphorylated pUC18 DNA was used as cloning vector. Library construction and screening was as above. Extensive restriction enzyme maps were made for the isolated *Pst*I inserts and relevant fragments isolated for sequence analysis.

DNA sequencing

Suitably sized DNA fragments were ligated into M13mp18/19 vectors (Yanisch-Perron *et al.*, 1985) followed by transformation of JM109. Single strand templates were sequenced with the dideoxynucleotide chain termination method (Sanger *et al.*, 1977) using universal primer, Klenow enzyme (Boehringer) and [α-^{35}S]dATPαS (600 Ci/mmol, Amersham). Sequencing reactions were electrophoresed on 0.4 mm thin 6% polyacrylamide−8 M urea gels. All exon and most intron sequences were determined at least twice on both strands.

Acknowledgements

We gratefully acknowledge the skilful technical assistance of Rita Feldmann and Beate Preitz. We are indebted to Eckhard Kube for his expert computer designed artwork. We thank Drs Volker Gerke and Mary Osborn for discussion. We also wish to thank Claudia Hake for photography and Elizabeth Crawford for preparation of the manuscript. The sequences reported above are available from the EMBL/GenBank under accession numbers X55947 to X55953.

References

Aebi,U., Cohn,J., Buhle,L. and Gerace,L. (1986) *Nature*, **323**, 560−564.
Alt,F.W., Bothwell,A.L.M., Knapp,M., Siden,E., Mather,E., Koshland,M. and Baltimore,D. (1980) *Cell*, **20**, 293−301.
Aviv,H. and Leder,P. (1972) *Proc. Natl. Acad. Sci. USA*, **69**, 1408−1412.
Bader,B.L., Magin,T.M., Hatzfeld,M. and Franke,W.W. (1986) *EMBO J.*, **5**. 1865−1878.
Bailey,J.M. and Davidson,N. (1976) *Anal. Biochem.*, **70**, 75−85.
Balcarek,J.M. and Cowan,N.J. (1985) *Nucleic Acids Res.*, **13**, 5527−5543.
Bartnik,E., Osborn,M. and Weber,K. (1985) *J. Cell Biol.*, **101**, 427−440.
Bartnik,E., Kossmagk-Stephan,K. and Weber,K. (1987a) *Eur. J. Cell. Biol.*, **44**, 219−228.
Bartnik,E., Kossmagk-Stephan,K., Osborn,M. and Weber,K. (1987b) *Eur. J. Cell. Biol.*, **43**, 329−338.
Beams,H.W., Tahmasian,T.N., Devine,R. and Anderson,E. (1957) *Exp. Cell Res.*, **13**, 200−204.
Birnstiel,M.L., Busslinger,M. and Strub,K. (1985) *Cell*, **41**, 349−359.
Brawerman,G. (1987) *Cell*, **48**, 5−6.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349−383.
Casey,J. and Davidson,N. (1977) *Nucleic Acids Res.*, **4**, 1539−1552.
Cerezuela,M.A. and Moreno Dias de la Espina,S. (1990) In Harris,J.E. (ed.), *Nuclear Structure and Function*. Plenum Press, New York, in press.
Dibb,N.J. and Newman,A.J. (1989) *EMBO J.*, **8**. 2015−2021.
Dodemont,H., Groenen,M., Jansen,L., Schoenmakers,J. and Bloemendal,H. (1985) *Biochim. Biophys. Acta*, **824**, 284−294.
Döring,V. and Stick,R. (1990) *EMBO J.*, **9**, 4073−4081.
Early,P., Rogers,J., Davis,M., Calame,K., Bond,M., Wall,R. and Hood,L. (1980) *Cell*, **20**, 313−319.
Fawcett,D.W. (1966) *Am. J. Anat.*, **119**, 129−146.
Fawcett,D.W. (1981) In *The Cell*. 2nd edn. W.B.Saunders Co., Philadelphia, p. 281.
Fisher,D.Z., Chaudhary,N. and Blobel,G. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 6450−6454.
Geisler,N. and Weber,K. (1982) *EMBO J.*, **1**. 1649−1656.
Georgatos,S.D., Maroulakou,I. and Blobel,G. (1989) *J. Cell Biol.*, **108**, 2069−2082.
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.
Gubler,U. and Hoffman,B.J. (1983) *Gene*, **25**, 263−269.
Hanahan,D. (1985) In Glover,D.M. (ed.), *DNA Cloning: A Practical Approach*. IRL Press, Oxford, Vol. I, pp. 109−135.
Hanahan,D. and Meselson,M. (1980) *Gene*, **10**, 63−67.
Hancock,J.F., Magee,A.I., Childs,J.E. and Marshall,C.J. (1989) *Cell*, **57**, 1167−1177.
Hargreaves,A.J., Goodbody,K.C. and Lloyd,C.W. (1989) *Biochem. J.*, **261**, 679−682.
Hennessy,S.W., Frazier,B.A., Kim,D.D., Deckwerth,T.L., Baumgartel, D.M., Rotwein,P. and Frazier,W.A. (1989) *J. Cell Biol.*, **108**, 729−736.
Holtz,D., Tanaka,R.A., Hartwig,J. and McKeon,F. (1989) *Cell*, **59**, 969−977.
Ichinose,Y., Morita,T., Zhang,F., Srimahasongcram,S., Tondella,M.L.C., Matsumoto,M., Nozaki,M. and Matsushiro,A. (1988) *Gene*, **70**, 85−95.
Johnson,L.D., Idler,W.W., Zhou,X.-M., Roop,D.R. and Steinert,P.M. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1896−1900.
Julien,J.-P., Grosveld,F., Yazdanbaksh,K., Flavell,D., Meijer,D. and Mushynski,W. (1987) *Biochim. Biophys. Acta*, **909**, 10−20.
Julien,J.-P., Côté,F., Beaudet,L., Sidky,M., Flavell,D., Grosveld,F. and Mushynski,W. (1988) *Gene*, **68**, 307−314.
Kozak,M. (1989) *J. Cell Biol.*, **108**, 229−241.
Krauss,S. and Franke,W.W. (1990) *Gene*, **86**, 241−249.
Krieg,T.M., Schafer,M.P., Cheng,C.K., Filpula,D., Flaherty,P., Steinert,P.M. and Roop,D.R. (1985) *J. Biol. Chem.*, **260**, 5867−5870.
Lees,J.F., Shneidman,P.S., Skuntz,S.F., Carden,M.J. and Lazzarini,R.A. (1988) *EMBO J.*, **7**, 1947−1955.
Leff,S.E., Rosenfeld,M.G. and Evans,R.M. (1986) *Annu. Rev. Biochem.*, **55**, 1091−1117.
Lendahl,U., Zimmerman,L.B. and McKay,R.D.G. (1990) *Cell*, **60**, 585−595.
Lewis,S.A. and Cowan,N.J. (1986) *Mol. Cell. Biol.*, **6**, 1529−1534.
Loewinger,L. and McKeon,F. (1988) *EMBO J.*, **7**, 2301−2309.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Marchuk,D., McCrohon,S. and Fuchs,E. (1984) *Cell*, **39**, 491−498.
McKeon,F.D., Kirschner,M.W. and Caput,D. (1986) *Nature*, **319**, 463−468.
Mercer,E.H. (1958) *Proc. R. Soc.B.*, **150**, 216−232.
Mitchell,P.J. and Tjian,R. (1989) *Science*, **245**, 371−378.
Myers,M.W., Lazzarini,R.A., Lee,V.M.-Y., Schlaepfer,W.W. and Nelson,D.L. (1987) *EMBO J.*, **6**, 1617−1626.
Osborn,M. and Weber,K. (1986) *Trends Biochem. Sci.*, **11**, 461−472.
Padgett,R.A., Grabowski,P.J., Konarska,M.M., Seiler,S. and Sharp,P.A. (1986) *Annu. Rev. Biochem.*, **55**, 1119−1150.
Palmiter,R.D. (1974) *Biochemistry*, **13**, 3606−3615.
Pappas,G.D. (1956) *J. Biophys. Biochem. Cytol.*, **2**, 431−435.
Peacock,S.L., McIver,C.M. and Monahan,J.J. (1981) *Biochim. Biophys. Acta*, **655**, 243−250.
Pollard,K.M., Chan,E.K.L., Grant,B.J., Sullivan,K.F., Tan,E.M. and Glass,C.A. (1990) *Mol. Cell. Biol.*, **10**, 2164−2175.
Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature*, **263**, 211−214.
Quax,W., Vree Egberts,W., Hendriks,W., Quax-Jeuken,Y. and Bloemendal,H. (1983) *Cell*, **35**, 215−223.
Quax,W., van den Broek,L., Vree Egberts,W., Ramaekers,F., and Bloemendal,H. (1985) *Cell*, **43**, 327−338.
Rigby,P.W.J., Dieckmann,M., Rhodes,C. and Berg,P. (1977) *J. Mol. Biol.*, **113**, 237−251.
Rogers,J., Early,P., Carter,C., Calame,K., Bond,M., Hood,L. and Wall,R. (1980) *Cell*, **20**, 303−312.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463−5467.
Shaw,G. and Kamen,R. (1986) *Cell*, **46**, 659−667.
Sippel,A.E. (1973) *Eur. J. Biochem.*, **37**, 31−40.
Southern,E.M. (1975) *J. Mol. Biol.*, **98**, 503−517.
Steinert,P.M. and Roop,D.R. (1988) *Annu. Rev. Biochem.*, **57**, 593−625.
Steinert,P.M. and Liem,R.K.H. (1990) *Cell*, **60**, 521−523.
Thomas,P.S. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 5201−5205.
Thompson,M.A. and Ziff,E.B. (1989) *Neuron*, **2**, 1043−1053.
Tyner,A.L., Eichman,M.J. and Fuchs,E. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4683−4687.
Vorburger,K., Kitten,G.T. and Nigg,E.A. (1989) *EMBO J.*, **8**, 4007−4013.
Weber,K., Plessmann,U., Dodemont,H. and Kossmagk-Stephan,K. (1988) *EMBO J.*, **7**, 2995−3001.
Weber,K., Plessmann,U. and Ulrich,W. (1989) *EMBO J.*, **8**, 3221−3227.
Wilson,T. and Treisman,R. (1988) *Nature*, **336**, 396−399.
Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) *Gene*, **33**, 103−119.

Received on July 16, 1990; revised on August 23, 1990