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; Grunenbaum et al., 1988), the nuclear lamins seem to constitute together with the cytoplasmic IF proteins a superfamilly, 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 Helix 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 originate 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 unfractuated 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 E3 and E4 each selected both mRNAs simultaneously although with different efficiency. Whereas hybrid selection with plasmid E3 (lane 2) was heavily biased towards the B-encoding mRNA, both mRNAs were retained by plasmid E4 (lane 3) to an extent comparable with their relative abundance in the original mRNA population (lane 4). Final evidence that the E3 and E4 plasmid represent the B and A mRNA sequence was obtained by sequencing the cDNA inserts. These are shown schematically in Figure 2B. Plasmids pSonIF52 E3 and pSonIF66 E4 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 E3B to prime the original mRNA. Clone pSonIF 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.](image)

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.
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) when 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.

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
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 PstI digested genomic DNA. Two fragments of 9.4 and 2.6 kb were detected by the pSonnIF PE-1 insert, whereas the subcloned E3A and E3B portions from pSonnIF52 E3 hybridized to 11.4 and 0.8 kb fragments, respectively. The entire pSonnIF66 E4 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 E3A probe. All seven PstI fragments were cloned separately. Their position within the overall genomic organization was established using appropriate subprobes in comparative Southern blot hybridization analyses of single and multiple digestives with BamHI, EcoRI, HindIII and PstI. An eighth 0.16 kb PstI fragment, which escaped detection by the E3B probe, was isolated on a 9.6 kb BamHI–HindIII fragment bridging the gap between the 2.6 and 0.8 kb PstI 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 sequences were contained within the series of eight PstI 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.

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

<table>
<thead>
<tr>
<th>Consensus</th>
<th>5′</th>
<th>3′</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNAG gtaagt</td>
<td>yyyy-yyyyyyyyynccag</td>
<td>GNNNN</td>
</tr>
<tr>
<td>GluLe</td>
<td>161</td>
<td>162</td>
</tr>
<tr>
<td>IA</td>
<td>GAGCT gtaag...</td>
<td>(22)</td>
</tr>
<tr>
<td>223</td>
<td>224</td>
<td></td>
</tr>
<tr>
<td>Ala</td>
<td>Asp</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>GTGCA gtaag...</td>
<td>(0.745)</td>
</tr>
<tr>
<td>255</td>
<td>256</td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>Leu</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>ACCAG gtaact...</td>
<td>(0.393)</td>
</tr>
<tr>
<td>308</td>
<td>309</td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>Leu</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>CACAG gtgasg...</td>
<td>(0.994)</td>
</tr>
<tr>
<td>350</td>
<td>351</td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>Phe</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>CCAAG gtaga...</td>
<td>(1.345)</td>
</tr>
<tr>
<td>424</td>
<td>425</td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>Gln</td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>TCTAG gtagt...</td>
<td>(1.4)</td>
</tr>
<tr>
<td>453</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>Ser</td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>ATCCA gtagt...</td>
<td>453</td>
</tr>
<tr>
<td>Ser</td>
<td>Trp</td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>ATCCA gtagt...</td>
<td>(22)</td>
</tr>
<tr>
<td>492</td>
<td>493</td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>Thr</td>
<td></td>
</tr>
<tr>
<td>VIII</td>
<td>CCAAG gtgact...</td>
<td>(1.6)</td>
</tr>
<tr>
<td>527</td>
<td>528</td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>Ile</td>
<td></td>
</tr>
<tr>
<td>IX</td>
<td>ACCAC gtaga...</td>
<td>(1.5)</td>
</tr>
<tr>
<td>565</td>
<td>566</td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>Glu</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>GAAAG gtagc...</td>
<td>(2.3)</td>
</tr>
</tbody>
</table>

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 4A). 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 polyproteins. Precise or estimated lengths (kb) of the introns are given in parentheses. Asterisks and circles refer to identically positioned introns for the Xenopus laevis LII gene (Döring and Stick, 1990) and the hamster vimentin gene (Quax et al., 1983), respectively.
The nucleotides differing only 4088 proteins (Weber middle by H.Dodemont, TGTGCCTAAT TCCAGATTT ACAGACAAGGCAGACCAAGATAAATGATGCTAGGAGCCTAAGATGAAGCACAGCAGCCAAGGCAGACAGGTAATGATGCTAGGAGCCTACGATGGTGACAT TTACACACACA CACTGGAGACTGTCTCTAT TCCACTGT TGACAAGACAAGATAAT ATGTC TCTGTGT TTAT AAACAAACAACAT TTTATGGACTTAATACAAGACAT ATAAT CAGGATATATTGAAATGTACAGAAAACTAAGCAACAGATACTCAAAATCAAAGCAGTGT TCTCATTTCAGTACGTCTGATACTGTAAACTGTCGAC---------------------------------------------------

GGGACCAA TTGCG -----------------------------------------------------------------------

GTGTTCTATAGCAACATAGGGCTACTGTAACT ATATAAATGATAGAGTAGTTATATGAGCATAGT AGTTTCTATCAACTCGCTGACACCGAGAATATCCAGATAGAAATCGTTCACAATTCTACACACT TTCCCAGGTAGGATAGTATGGGTATGTATAAACACATTGTAGTCCCTGCCGTAT CACTTAATGCGAGGAAACCCAAGCAAATGTTTCATGTGCCTACATTCATGTGGCTGCTTCACTCGTGTAGATAGAGATAACCTGCTGCCTAATGTGTTCCATGTATACCAATCTGATATGATAATTAAGAAACTTCTAATTTACTGCAAATAATGCAGATACTTTGCAAATAATACTAATAACTAATAACATATAGTACTGACAGACAAAAATCGTAAGCATCTAATTCATCACATGCTAAGTAGTCAAACTGTAAATAGTCAATCACATTACAGCTGAAGAACCTTATG

---

exon genomic sequences reading from the TAG and TAA stop codon four out two positions were found cDNAs. Sizes of translational initiation TATA-like sequences, positions with the ATG codon shows located in the 3'-ends of the cDNAs. 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 slashed lines. The genomic sequence of the H.pomatia cDNA was represented in cDNA as CTTTTT and AGTC, 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 repeat sequence in genomic DNA (located in exon 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 GGCAGCACGAGG, considered to be a favourable context for translational initiation (Kozak, 1989). The predicted H.aspera 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

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 nucleotide sequences. 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 slashed lines. The genomic sequence of the H.pomatia cDNA was represented in cDNA as CTTTTT and AGTC, 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 repeat sequence in genomic DNA (located in exon 7) was found in different copy numbers in several cDNAs.
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 PstI 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 VIF IF genes and the invertebrate nn IF gene (Figures 7 and 8). The strict conservation of the first six intron positions (see also Table 1), 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 rod 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 identicaly 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* nIF, hamster vimentin and *Xenopus* lamin LIII gene organizations. Predicted protein sequences of the *Helix* nIF 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 lamin 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* nIF 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 coi 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* nIF and the lamin gene. Note that introns 8–10 of the nIF gene correspond, both in position and phase (see Table I), to lamin introns 7–9. The position of the unique intron 7 of the nIF 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 nIF gene, separates the last exon, which carries the CaaX motif (asterisks) from exon 9, whose 5'-sequence is aligned with the C-terminal end of the Helix protein. Sequence alignment of seven lamin proteins including that of *Drosophila* (Polland *et al.*, 1990) shows 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* nIF 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.*, 1988) the corresponding regions of this invertebrate IF protein show no length variability versus lamin LIII.

**Discussion**

**Characteristics of expression of the non-neuronal IF gene**

We characterized the gene which encodes the nIF 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 nIF type in a variety of tissues (Bartnik *et al.*, 1987b). The nIF 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 retained 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...
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 laminins 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 laminin genes as was postulated earlier (Osborn and Weber, 1986; Bartnik et al., 1987a; Myers et al., 1987; Steinert and Roop, 1986).
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 Dörring 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 Días 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 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 lamin 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 Dörring 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 Biomaries 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 4°C in 5 ml of 20 mM Tris·HCl (pH 8.0) containing 1 mM EDTA, 0.5 mM DTT, 50 mM MgCl₂, 20% glycerol, and 20 μg/ml each of RNase A and RNase T₁. The homogenate was centrifuged at 10,000 × g for 10 min. The precipitate was washed twice in 1 ml of 20 mM Tris·HCl, 1 mM EDTA, 0.5 M DTT, and 20% glycerol and resuspended in 0.5 ml of 20 mM Tris·HCl, 1 mM EDTA, 0.5 M DTT, and 20% glycerol. The RNA was precipitated with an equal volume of 95% ethanol at -20°C for 1 h. The RNA was washed twice with 70% ethanol, dissolved in 1 ml of 20 mM Tris·HCl, 1 mM EDTA, and 0.5 M DTT, and stored at -80°C.

Identification of mRNA encoding cDNA

Poly(A)-containing polysomal RNA was electrophoresed in a 1.5% agarose gel in the presence of 10 mM methymercury hydroxide (Bailey and Davidson, 1976). The gel was stained with ethidium bromide and examined under ultraviolet light. The RNA bands were excised and eluted with sterile water. The poly(A) content of the RNA was determined by the method of Aviv and Leder (1972) using two cycles of affinity chromatography to oligo(dT)-cellulose (Pharmacia).

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 micrograms of this RNA was polyadenylated in vitro at 37°C for 1 h in a 40 μl reaction mixture containing 50 mM Tris·HCl, pH 8.0, 37°C, 250 mM NaCl, 10 mM MgCl₂, 10 mM DTT, 500 μg/ml BSA, 100 μM ATP, 0.5 μCi [α-32P]ATP (300 Ci/mmol, Amersham) and 1 U of polyadenylate nucleotidyltransferase (Sippel, 1973). The reaction was stopped by addition of EDTA, Na-sarcosyl and protease 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 EDTA, 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) tract of 80 residues.

cDNA synthesis

Oligo(dT)₃ 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 (Guberl 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, 5 mM MgCl₂, 10 mM DTT, 100 μg/ml oligo(dT)₁₂₋₁₈, 1 mM of each dNTP, 10 μCi [α-32P]dCTP (3000 Ci/mmol) and 50 U of reverse transcriptase (Amerham). 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 (NH₄)₂SO₄, 10 mM MgCl₂, 10 mM DTT, 100 μg/ml BSA, 100 μM of all four dNTPs, 2.5 μCi [α-32P]dCTP and 150 μM dNAD. The enzyme (E. coli, E. coli, 0.025 unit) was added and the reaction was incubated sequentially at 12°C and 18°C for 1 h each. The reaction was stopped by addition of 50 μl of 5× hybridization buffer (100 mM NaPO₄, pH 6.4, 400 mM NaCl, 5 mM EDTA, 0.2% Na-sarcosyl, 100 μg/ml yeast RNA and 200 μg/ml tRNA). 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 liquid and recovery of hybrid released mRNAs were as reported (Dedmon et al., 1985). The mRNAs were identified by gel electrophoretic analysis of immunoprecipitated in vitro translation products.

Positive hybrid selected translation

Specific mRNAs were hybrid selected from total poly(A)⁺ RNA using 50 μg denatured plasmid DNA bound to nitrocellulose filters in 50 μl reaction volumes containing 65% formamide, 50 mM Na-Pipes, pH 6.4, 400 mM NaCl, 5 mM EDTA, 0.2% Na-sarcosyl, 100 μg/ml yeast RNA and 200 μg/ml tRNA. 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 liquid and recovery of hybrid released mRNAs were as reported (Dedmon et al., 1985). The mRNAs were identified by gel electrophoretic analysis of immunoprecipitated in vitro translation products.

Transfer of alkali-treated DNA and glyoxylated RNA from agarose gels to nitrocellulose membranes was as described (Southern, 1975; Thomas, 1978). Filters were prehybridized in 2× SSC, 50 μg/ml salmon sperm DNA, and 50 μg/ml denatured salmon sperm DNA for 30 min at 42°C. For hybridization this mixture was replaced by a fresh solution of the same composition except that it contained 20 μg/ml poly(A)⁺ RNA, 1× 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 μg/ml poly(A)⁺ RNA, 1× Denhardt’s and 5 ng/ml bovine transfected cDNA insert (see above). After 18–24 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.
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 cosmids libraries. Instead, several partial genomic DNA libraries were constructed in parallel, each enriched (50- to 200-fold) for a PstI fragment hybridizing to selected cDNA subprobes (see text). Gel purified, PstI 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 PstI 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-35S]dATP/s500 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 Preitiz. 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/GentBank under accession numbers X55947 to X55953.

**References**