# Eight genes and alternative RNA processing pathways generate an unexpectedly large diversity of cytoplasmic intermediate filament proteins in the nematode *Caenorhabditis elegans*

**Huub Dodemont, Dieter Riemer, Neil Ledger and Klaus Weber**

Max Planck Institute for Biophysical Chemistry, Department of Biochemistry, PO Box 2841, D-37018 Goettingen, Germany

Communicated by K.Weber

**Cytoplasmic intermediate filament (IF) proteins of *Caenorhabditis elegans* are encoded by a dispersed multigene family comprising at least eight genes which map to three linkage groups. Exon sequences and intron patterns define three distinct subfamilies. While all eight IF genes display the long coil 1b subdomain of nuclear lamins, only six genes ($a_1-a_4$, $b_1$ and $b_2$) retain a lamin-like tail domain. Two genes ($c_1$ and $c_2$) have acquired entirely novel tail domains. The overall sequence identity of the rod domains is only 29%. The gene structures show a strong drift in number and positions of introns, none of which are common to all genes. Individual genes share only one to four intron locations with the *Helix aspersa* IF gene, but all eight nematode genes together account for nine of the 10 introns of the gastropod gene. All *C.elegans* IF genes are transcribed and all except gene $c_2$ produce *trans*-spliced mRNAs. Alternatively spliced mRNAs arise from genes $a_1$, $b_2$ and $c_2$ through several mechanisms acting at the transcriptional and post-transcriptional levels. These involve the alternative use of distinct promoters, polyadenylation sequences and both *cis* and *trans* RNA splice sites. The resulting sequence variations are restricted to the non-helical end domains. Minimally 12 distinct IF proteins are encoded by the various mRNAs. Different abundances in mixed-stage nematode populations suggest cell type- and/or stage-specific expression of individual mRNAs.**

*Key words:* evolution/exon–intron structure/intermediate filaments/lamin homology/*trans* mRNA splicing

## Introduction

The multigene family of structural proteins of the cytoplasmic intermediate filaments (IF) comprises close to 50 different members in a mammal and can be subdivided into five types. The largest and most complex groups are provided by the type I and II keratins. Their products lead to the obligatory heteropolymeric keratin filaments of epithelia, which are built from double-stranded coiled coils containing one keratin I and one keratin II polypeptide. Vimentin and its close relatives cover the type III IF while neurofilament proteins encompass the type IV IF. The nuclear lamins, the structural proteins of the lamina, are usually considered as type V (for reviews see Steinert and Roop, 1988; Fuchs and Weber, 1994). Aberrant filament assembly due to point mutations in one or the other epidermal keratin is the basis of human skin blistering diseases such as epidermolysis bullosa simplex and epidermolytic hyperkeratosis (see for instance Coulombe *et al.*, 1991; Lane *et al.*, 1992; Rothnagel *et al.*, 1992). Outside the vertebrates molecular knowledge of cytoplasmic IF is very sparse and centers essentially on a few IF proteins and genes from several molluscs such as *Helix*, *Aplysia*, *Loligo* and *Octopus* (Weber *et al.*, 1988; Dodemont *et al.*, 1990; Riemer *et al.*, 1991; Szaro *et al.*, 1991; Way *et al.*, 1992; Tomarev *et al.*, 1993) and the large nematode *Ascaris lumbricoides* (Weber *et al.*, 1989). The cytoplasmic IF proteins of these invertebrates are more closely related to nuclear lamins than are the IF proteins of vertebrates. All of them have the lamin-like length of the coil 1b domain and nearly all of them contain a lamin-like tail domain. Thus it is thought that the archetypal cytoplasmic IF protein arose in eukaryotic evolution from a mutated lamin gene which lost two signal sequences related to lamin functionality i.e. the nuclear localization signal and the CaaX box (Weber *et al.*, 1989; Dodemont *et al.*, 1990; Döring and Stick, 1990).

Nematodes may offer a unique possibility for functional studies of IF centering on the nature of the epidermal structures involved in transmitting the tension generated by the muscles to the body surface. Electron microscopical studies on *A.lumbricoides* have shown that the thin epidermis which faces the muscle contains prominent bundles of desmosome- and hemidesmosome-linked tonofilaments that may function in force coupling (Rosenbluth, 1967; Bartnik *et al.*, 1986; Bartnik and Weber, 1987). These tonofilament arrays resemble IF by ultrastructural criteria, and immunofluorescence microscopy shows that they react with the monoclonal antibody IFA (Bartnik *et al.*, 1985, 1986), which recognizes many IF proteins (Pruss *et al.*, 1981). Two *Ascaris* proteins have been purified and shown to form IF *in vitro*. The amino acid sequences of these proteins show the structural features typical of invertebrate IF proteins (Weber *et al.*, 1989). In the case of the small nematode *Caenorhabditis elegans*, IF characterization is less advanced. While electron microscopy has documented IF in all cell types of *Ascaris*, ultrastructurally convincing *C.elegans* IF have been readily documented only for the marginal cells of the pharynx (Albertson and Thomson, 1976; Bartnik *et al.*, 1986; Francis and Waterston, 1991). Immunofluorescence microscopy with antibody IFA provides filamentous decoration of practically all cells of *Ascaris* while only certain *C.elegans* cells show such decoration (Bartnik *et al.*, 1986; Francis and Waterston, 1991). The latter observation may reflect the presence of several different IF proteins in *C.elegans* since small conservative replacements in the IFA epitope lead to a loss of IFA antigenicity of ultrastructurally normal IF in other invertebrates (Bartnik *et al.*, 1987; Bartnik and Weber, 1989; Riemer *et al.*, 1991). In addition, recent electron microscopic studies have made a good case for the existence of IF-related filaments also in epidermis and muscle of *C.elegans*. Thus in *C.elegans* IF may be involved in the transmission of tension from the muscle cell to the cuticle (Francis and Waterston, 1991).

A molecular and genetic approach to *C.elegans* IF genes seems promising for the following reasons. First, epidermal IF may participate in transmitting the tension from the muscle to the cuticle and thus participate in the locomotion of the nematode (see above). Second, *C.elegans* offers the possibility of studying physiological functions of cytoskeletal proteins by a combination of molecular and genetic approaches (see for instance Barstead and Waterston, 1991). Third, the genome of *C.elegans* may be completely sequenced in the next few years (Sulston *et al.*, 1992). Here we provide a molecular analysis of IF-encoding sequences of *C.elegans*. We document a multigene family of at least eight members. Differential RNA splicing pathways increase the total number of IF proteins to 12 distinct species.

## Results

### Assessment of IF gene complexity in C.elegans

Sequences defining the coil 1a subdomain of non-neuronal IF proteins from the mollusc *Helix aspersa* (Dodemont *et al.*, 1990) have been used to isolate by cross-hybridization novel IF-specific cDNAs from a distantly related metazoan species (Riemer *et al.*, 1992). By the same approach we have obtained and characterized four complete IF-encoding cDNAs from the nematode *A.lumbricoides* (H.Dodemont, N.Ledger, D.Riemer and K.Weber, unpublished results; see Materials and methods), two of which matched the major IF protein sequences A and B (Weber *et al.*, 1989). DNA fragments comprising the *Ascaris* coil 1a regions were subcloned, mixed in equimolar amounts and hybridized under non-stringent conditions to Southern blot transfers of *C.elegans* genomic DNA digested with various restriction enzymes. Following a low stringency wash (5 × SSC, 55°C) every digest displayed seven to 10 fragments hybridizing at widely varying intensities (Figure 1A). Many of the weaker bands were no longer detected after washing at intermediate stringency (0.5 × SSC, 55°C) and only the most strongly hybridizing fragments (one or two per digest) persisted through the final wash at high stringency (0.1 × SSC, 60°C) (data not shown). In a parallel experiment a duplicate blot of the one shown in Figure 1A was hybridized with DNA sequences which generally are far less conserved among different IF genes than the coil 1a regions. These hybridization probes comprised the entire tail domains from two cloned *Ascaris* IF cDNAs (see Materials and methods). Again, multiple band patterns were obtained for every digest after washing at low stringency (Figure 1B). Surprisingly, nearly all of the hybridizing fragments were still present following the medium stringency wash (not shown), in contrast to the results obtained with the coil la probes. Together, the two largely non-overlapping sets of multiple hybridizing bands indicate that, first, IF sequences in *C.elegans* are organized as a multigene family, and second, among nematodes not only the coil la but also the tail domains of at least some IF genes show substantial sequence conservation.

### Isolation of putative IF genes

A genomic library of partially *Sau*3A-digested *C.elegans* DNA in the phage vector λ2001 (Karn *et al.*, 1984; Coulson *et al.*, 1986) and a series of new plasmid libraries, each representing highly enriched subgenomic fractions, were searched for *C.elegans* IF genes. For the initial screening of the λ genomic library the same conditions were used as for the Southern blots in Figure 1. Sequential plaque cross-hybridization with the heterologous *Ascaris* coil la and tail domain DNA probes provided two groups of genomic clones. Analysis of the first group of clones which hybridized with both *Ascaris* probes, identified three distinct complete IF genes. Based on sequence and structural criteria, these genes were designated CelIF $a_1$, $a_2$ and $b_1$. Only one novel IF gene was isolated from the second group of clones, which hybridized exclusively to the coil 1a probes. This gene, CelIF $c_2$, showed a tail domain which totally diverged from the corresponding sequences of the other three *C.elegans* IF genes and the *Ascaris* probes (see below). Rescreening of the λ genomic library with all four *C.elegans* coil 1a sequences under low stringency conditions did not reveal additional candidate IF genes. Hybridization of the original Southern blots (Figure 1A) with the same probes largely reproduced the multi-band patterns obtained previously with the *Ascaris* coil la sequences. Three coil 1a hybridizing DNA fragments, which were not derived from the four genes already isolated, were cloned from libraries representing the appropriate gel-purified size fractions. Subsequent genomic walking procedures and additional subgenomic libraries provided the genes CelIF $a_3$, $a_4$ and $c_1$. An eighth gene, CelIF $b_2$, was isolated by PCR (Saiki *et al.*, 1988) as a series of three overlapping genomic fragments using specific primers designed from the corresponding cDNA sequence (see Materials and methods).

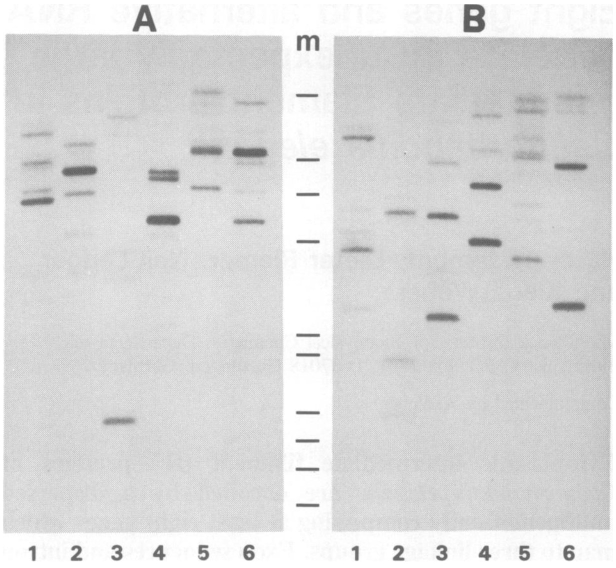

**Fig. 1.** Genomic representation of putative IF sequences in *C.elegans*. Duplicate digests (2 μg each) of genomic DNA with *Eco*RI, *Hin*dIII, *Sac*I, *Sal*I, *Xba*I and *Xho*I were electrophoresed on a single 0.5% agarose gel (lanes 1−6 respectively). Following transfer to a nitrocellulose membrane, DNA fragments were hybridized at low stringency to coil 1a (A) and tail domain (B) probes of *Ascaris* IF cDNAs (see Materials and methods). An 18 h autoradiograph is shown after the first wash in 5 × SSC/0.1% SDS at 55°C. Lane m gives from top to bottom the positions of DNA size standards (λ/*Hin*dIII and φX174/*Hae*III digests): 23.1, 9.4, 6.6, 4.4, 2.3, 2.0, 1.4, 1.1, 0.9 and 0.6 kb.

| C. eleg. IF gene | b1 | b2 | a1 | a2 | a3 | a4 | c1 | c2 |
|---|---|---|---|---|---|---|---|---|
| Gene location | II<br>Y 52 E8<br>or<br>I<br>Y 65 E10 | II<br>Y 52 E8<br>or<br>I<br>Y 65 E10 | X<br>Y 11 D12<br>Y 43 G6<br>Y 43 H4 | X<br>Y 43 C8<br>Y 44 B5<br>Y 49 E3<br>Y 50 A2<br>V<br>Y 44 H11 | X<br>Y 43 C8<br>Y 44 B5<br>Y 49 E3<br>Y 50 A2<br>V<br>Y 44 H11 | X<br>Y 12 F1<br>Y 50 C8 | V<br>Y 51 D5 | X<br>Y 42 G5<br>Y 51 E2<br>Y 55 B3 |
| Gene size (kb) | 3.0 | 4.4 | 4.3 | 3.4 | 2.9 | 3.7 | 2.4 | 3.6 |
| Intron number | 6 (4) | 8 (4) | 8 (2) | 5 (4) | 5 (4) | 9 (4) | 4 (1) | 10 (1) |
| Intron size (bp) | 45 -<br>457 | 50 -<br>~ 1380 | 47 -<br>1997 | 55 -<br>943 | 44 -<br>386 | 43 -<br>516 | 43 -<br>298 | 47 -<br>824 |
| mRNA size (nt) | 1946 | > 1797 (H)<br>> 1746 (L) | 2004 (H)<br>1953 (M)<br>1867 (L) | 1953 | 2079 | 2060 | 1702 | 1985 (H)<br>1975 (L) |
| 5' UTR size (nt) | 38 | 32 | 86 (H,M)<br>24 (L) | 27 | 33 | 25 | 24 | 58 |
| 3' UTR size (nt) | 234 | > 136 | 142 | 183 | 303 | 310 | 178 | 61 (H)<br>411 (L) |
| Trans-splicing | SL1 | SL1 | SL1 (H,M,L)<br>SL2 (L) | SL1 | SL1 | SL1 | SL1 | – |
| EST | – | cm 01g8<br>cm 06d5<br>cm 14g10 | – | cm 08g8 | – | – | cm 11f10 | cm 04h9 |

**Fig. 2.** Characteristics of *C.elegans* IF genes and their mRNAs. IF gene designations are shown in the top line. Gene locations on linkage groups (II or I, V and X) are defined by the YAC clones listed. Grouped YAC clones comprise single contigs. Note that the $a_2$/$a_3$ genes map to the same contig represented by two copies on separate chromosomes (V and X). The $b_1$/$b_2$ genes are covered by a single YAC clone whose identity has not been firmly resolved (see Results). Sizes of the *trans*-splicing genes are given as the distances between the 5′-most *trans*-splice acceptor and 3′-most poly(A) addition sites. In the case of the $c_2$ gene, which does not exert *trans*-splicing, the 5′-boundary has been assigned to the presumptive promoter element TATAAAA (Breathnach and Chambon, 1981) that occurs 23 bp upstream to the 5′-end of the largest cDNA. Intron numbers specify total introns present in each IF gene. Notations between parentheses refer to the number of introns with precisely the same locations as the corresponding introns in the gene encoding non-neuronal IF proteins of the mollusc *H.aspersa* (Dodemont *et al.*, 1990; see Figure 10). Together the eight *C.elegans* IF genes contain 55 introns, all of which interrupt coding sequences. Introns vary widely in size (from 43 to 1997 bp); 34 introns are smaller than 100 bp and 19 of these have a length of 50 bp or less. The first intron of the $b_2$ gene is the only intron which has not been completely sequenced. Its size (~1380 bp) was determined by gel electrophoresis. All intron border regions follow closely the consensus sequences (Blumenthal and Thomas, 1988; Fields, 1990) represented as $5'-A^{61}-G^{63}/g^{100}-t^{100}-a^{55}-a^{59}-g^{82}-t^{61}-3'$ and $5'-t^{84}-t^{96}-t^{67}-c^{86}-a^{100}-g^{100}/$N-3′, respectively for the 5′- and 3′-splice sites (lower case letters specify intron sequences; numbers indicate percentage of introns from IF genes with the designated base at the given position). Sizes are listed for the full-length IF mRNAs excluding poly(A) tails, and for their individual 5′- and 3′-untranslated regions (UTRs). Complete 3′-UTRs have been characterized for all mRNAs except for the $b_2$ mRNA whose 3′-UTR is at least 136 nt. Where appropriate, 5′-UTR sizes include the 22 nt of the *trans*-spliced leader sequences SL1 (Krause and Hirsh, 1987) or SL2 (Huang and Hirsh, 1989). H, M and L refer to high, medium or low molecular weight mRNA variants arising by differential RNA splicing pathways. Note that the $a_1$-L RNA transcript may be alternatively spliced *in trans* to either SL1 or SL2 whereas $a_1$-H and $a_1$-M RNAs *trans*-splice exclusively to SL1 (see Results). The expressed sequence tag (EST) designations are reported by Waterston *et al.* (1992).

### *IF gene analysis and exon prediction*

Genomic DNA fragments and subclones were characterized by high resolution restriction enzyme mapping, followed by Southern blot hybridization to localize the coil 1a encoding regions. The sequences of all eight genes and their flanking regions were determined on both strands by first analyzing the coil 1a domain and then extending in both directions. Sequences which total 35 kb were obtained. The sizes of the genes varied from 2.4 to 4.4 kb (Figure 2). Like many other *C.elegans* genes (see Sulston *et al.*, 1992) the IF gene sequences proved to be particularly amenable for prediction of putative coding exons. Exon prediction was done for all genes except $b_2$. It relied heavily on the substantial homology with the two established *Ascaris* IF protein sequences (Weber *et al.*, 1989) which was anticipated from the hybridization data. Interpretation of the gene sequences was further facilitated by the generally small size of *C.elegans* introns and their well defined 5′- and 3′-splice site consensus sequences (Blumenthal and Thomas, 1988; Fields, 1990). These criteria allowed direct identification of the protein coding sequences of all seven rod domains, of the tail domains, except those of $c_1$ and $c_2$, and of the C-terminal portions of the head domains of $a_1$, $a_2$, $a_3$, $a_4$ and $b_1$. Most other *C.elegans* IF protein coding sequences could be predicted as long as their reading frames extended contiguously beyond previously assessed coding regions. This applied to the highly divergent tail domains of $c_1$ and $c_2$ and the complete head domains of $a_2$, $a_3$, $c_1$ and $c_2$.

Typical examples of *C.elegans* IF genes whose coding exons were fully predicted are the $a_2$ and $a_3$ genes, which are shown in Figure 3. The two genes are highly related; they show sequence identity values for the entire coding regions of 84 and 87% at the nucleic acid and protein level respectively. Straightforward identification of all coding exons covering the rod and tail domains was supported by an 80% protein sequence similarity with the corresponding regions of *Ascaris* IF protein A (Weber *et al.*, 1989). An even higher similarity ($a_2$ versus AscIF A: 89%; $a_3$ versus AscIF A: 82%) is found in the C-terminal portions (28 residues) of the respective head domains. The lack of pronounced sequence similarity with the *Ascaris* IF protein in the preceding part of the head domains indicates sequence divergence and is not due to intron sequences since the corresponding $a_2$ and $a_3$ nucleic acid sequences are devoid of typical 3′- splice sites. Instead, the reading frames can be followed in the 5′-direction towards possible ATG translational start codons located immediately downstream from putative splice acceptor sites. Similar predictions for the full head domains were not possible for $a_1$, $a_4$ and $b_1$. Here the consensus splice acceptor sites 5′ to the sequences encoding the C-terminal regions of the head domains implied separate exons for the N-terminal sequences. These were

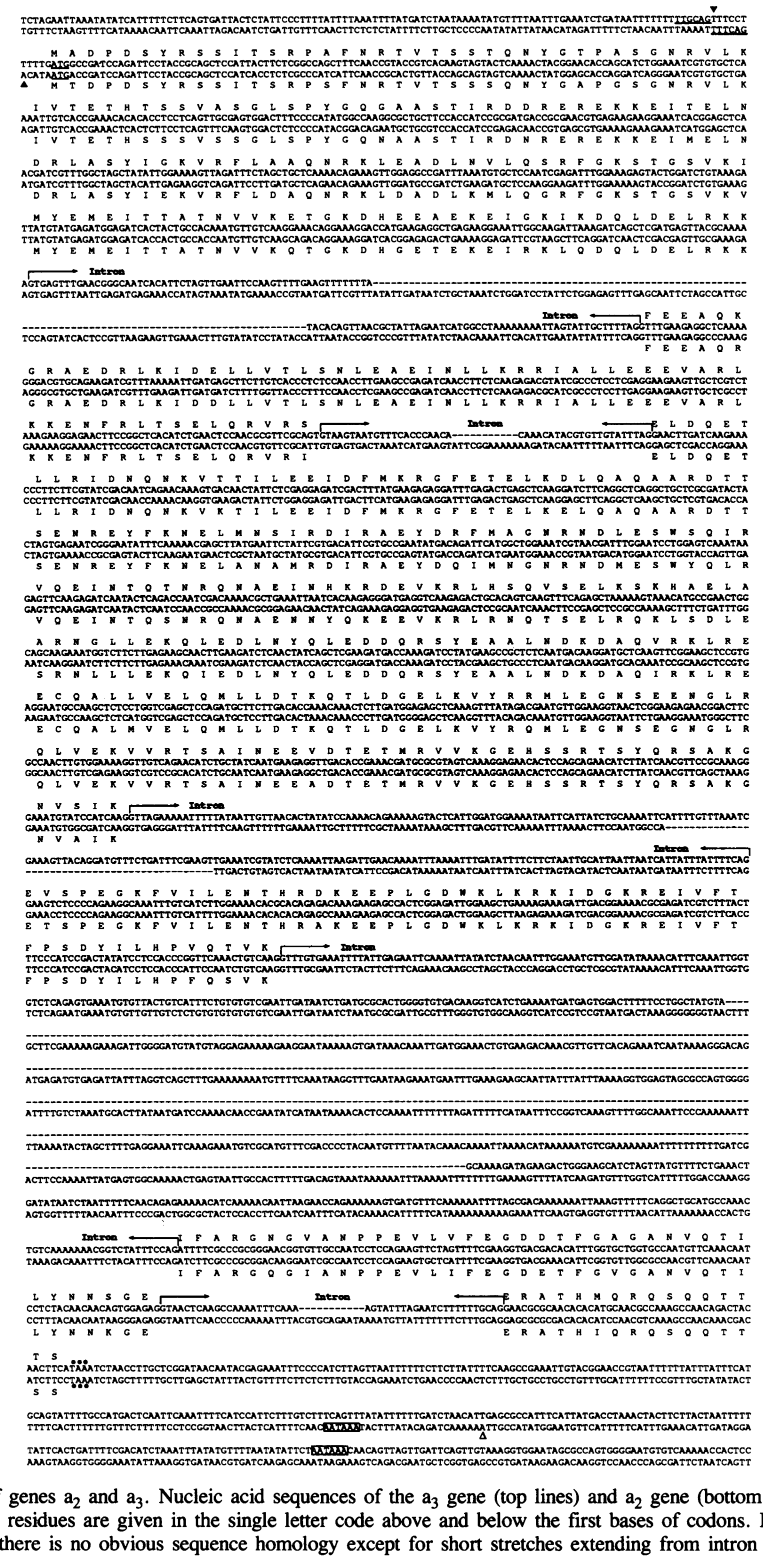

Fig. 3. Sequence alignment of genes $a_2$ and $a_3$. Nucleic acid sequences of the $a_3$ gene (top lines) and $a_2$ gene (bottom lines) are shown with 120 nt per line. Predicted amino acid residues are given in the single letter code above and below the first bases of codons. Introns are marked by arrows. Outside the coding sequences there is no obvious sequence homology except for short stretches extending from intron borders and an internal region near the 5′-end of the fourth intron. Dashes in intron sequences allow for compensation of intron size differences. Both ATG start codons (underlined) are preceded by consensus splice acceptor sites (underlined) for the SL1 leader sequence (Krause and Hirsh, 1987) which splices *in trans* to the nucleotides marked by closed arrowheads. The TAA stop codons are marked by dots and the canonical AATAAA polyadenylation signals (Proudfoot and Brownlee, 1976) are framed. The poly(A) addition site for $a_2$ RNA transcripts is marked by an open arrowhead. Such a site was not identified for $a_3$ RNA transcripts. Additional putative polyadenylation sequences (Birnstiel *et al.*, 1985; Spieth *et al.*, 1993) are represented by AATATA (10 nt upstream of AATAAA) for $a_3$ RNA and by ATTAAA and AATAAG (89 and 117 nt downstream of AATAAA, respectively) for $a_2$ RNA. The AATAAA polyadenylation signal of the $a_3$ gene occurs 1527 bp upstream to the translational start site of the $a_2$ gene (see Results).

the only obvious exons that could not be inferred directly from the gene sequences and thus required cDNA information (see below). The 45 exons predicted for the seven CelIF genes $a_1$, $a_2$, $a_3$, $a_4$, $b_1$, $c_1$ and $c_2$ together account for 95% of their total protein coding sequence potential.

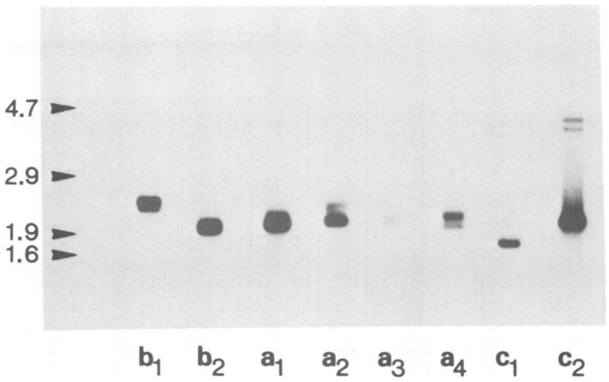

**Fig. 4.** Northern analysis of *C.elegans* IF mRNAs. Eight aliquots (5 $\mu$g each) of glyoxylated total poly(A)$^+$ RNA were electrophoresed in a single 1.5% agarose gel and blotted onto a nitrocellulose membrane. Individual filter strips were hybridized to the IF-specific DNA probes indicated (see Materials and methods) followed by washing to high stringency (0.1 × SSC/0.1% SDS at 60°C). The composite autoradiograph is shown after 18 h exposure. Positions of RNA size standards (rat 18S and 28S rRNA, *E.coli* 16S and 23S rRNA) are given.

### *Mapping of IF genes*

Physical map positions for the IF genes were obtained by hybridization of 'polytene' YAC grids (Coulson *et al.*, 1991; Waterston *et al.*, 1992) with gene-specific probes, followed by chromosomal localization of positive clones using the ACEDB genome database (Sulston *et al.*, 1992; Waterston *et al.*, 1992). The *C.elegans* IF multigene family is dispersed since the eight genes mapped to at least three linkage groups II (I), V and X (Figure 2). All genes were confined to single loci except for the $a_2$ and $a_3$ genes, each of which was assigned to the same two distinct loci on different chromosomes. The latter observation suggests that the two genes are represented as a cluster that occurs twice in the *C.elegans* genome. One copy was isolated here. The sequences of the $a_2$ and $a_3$ genes comprise a single contig revealing a tightly linked tandem of the two genes in the same orientation with the polyadenylation signal of $a_3$ 1527 bp upstream to the translational initiation codon of $a_2$. CelIF $b_1$ and $b_2$ form another gene pair. Both hybridized to a single YAC clone which, however, could not be unambiguously identified (Figure 2) due to poor resolution of the hybridization signal on the grid filter. Clustering of $b_1$ and $b_2$ is probably less tight than for the $a_2$/$a_3$ genes as no overlaps were found among the available sequence data.

### *Characterization and expression of IF mRNAs*

Northern analysis shows that all genes are transcribed (see below). Corresponding cDNAs were isolated from a large plasmid cDNA library comprising 400 000 primary clones representing total cellular poly(A)$^+$ RNA from mixed-stage worm populations. Library screening was by hybridization under stringent conditions using DNA probes from previously identified 5'- coding sequences of the genes. Full-length or near complete cDNAs were obtained for all genes except $a_2$ and $a_3$. Additional cDNA sequences were generated by reverse transcription (RT)-PCR. Sequence analysis confirmed all exons previously predicted at the gene level (see above) and allowed identification of the yet unresolved head domain sequences of $a_1$, $a_4$ and $b_1$. Four different cDNAs were identified for the $a_1$ gene and two distinct cDNAs were obtained for the $c_2$ gene. All six cDNAs represent alternative RNA variants which arise by differential RNA splicing pathways (see below). In contrast, the several independent cDNAs isolated for each of the other IF genes were all derived from single RNA transcripts.

In a survey of moderately to abundantly expressed genes of *C.elegans*, Waterston *et al.* (1992) recently identified a set of six short cDNA sequences putatively encoding IF proteins. Three of these expressed sequence tags (ESTs) matched the *C.elegans* IF sequences $a_2$, $c_1$ and $c_2$ characterized here, whereas the other three represented novel IF sequences (Figure 2). The 5'-portion from one of the latter ESTs (cm0lg8) was recovered from total cellular mRNA template by RT-PCR as a 186 bp product. This was used as a screening probe to isolate by hybridization the corresponding complete cDNA from the plasmid library. Several overlapping cDNAs were obtained that represented two alternatively spliced variants from a single new gene designated CelIF $b_2$ (see below). All three novel EST sequences are comprised by the two different $b_2$ cDNAs.

In total 13 different *C.elegans* IF-specific cDNAs have been characterized. Complete 5'-untranslated sequences varying in size from 24 to 86 nt (Figure 2) were obtained for all of them (see below). Nine cDNAs contained remnants of poly(A) tails whereas the other four terminated near putative polyadenylation signals. Polyadenylation signals either were of the consensus type (Proudfoot and Brownlee, 1976; Birnstiel *et al.*, 1985) or were represented by unusual sequence motifs similar to those found in other *C.elegans* genes (Riemer *et al.*, 1993; Spieth *et al.*, 1993). For all cDNAs the onsets of the single large open reading frames were assigned to the first ATG codons. All of these are part of sequence contexts considered favorable for translational initiation as defined originally for vertebrate mRNAs (Kozak, 1991). Twelve ATG start codons are preceded by a purine base in the −3 position while the single ATG that displays here a pyrimidine ($a_3$ gene; see Figure 3) is flanked by the obligatory guanosine in the +4 position. The distinct base compositions around the initiation codons in conjunction with the sizes and/or sequences of the 5'-non-coding regions may reflect different translational activities of the IF mRNAs.

The molecular sizes predicted for the IF mRNAs assuming average poly(A) tails of 100 adenylate residues range from ~1800 to ~2200 nt, which were essentially confirmed by Northern analysis (Figure 4). All but the $c_1$ RNA transcript appeared as broad bands or discrete doublets. Size heterogeneity may arise from several RNA splice variants for a particular transcript (like $a_1$, $b_2$ and $c_2$, see below) or may result from differently sized 3'-untranslated sequences due to utilization of alternative polyadenylation sequences. The nature of the large minor transcripts detected by the $c_2$-specific probe is unknown (see below). Judged from the Northern blots all IF mRNAs are among the moderately expressed sequences in mixed-stage nematode populations. Relative hybridization signals vary over an estimated 50-fold range from the $a_1$ mRNA as the most prominent transcript to the $a_3$ mRNA as a rather minor species. Additional major mRNAs are represented by $b_1$, $b_2$ and $c_2$, whereas $a_2$, $a_4$ and $c_1$ are expressed at intermediate levels.

### *Cytoplasmic IF protein sequences*

All eight proteins, as well as the additional variants, display the IF-specific tripartate structural organization based on a

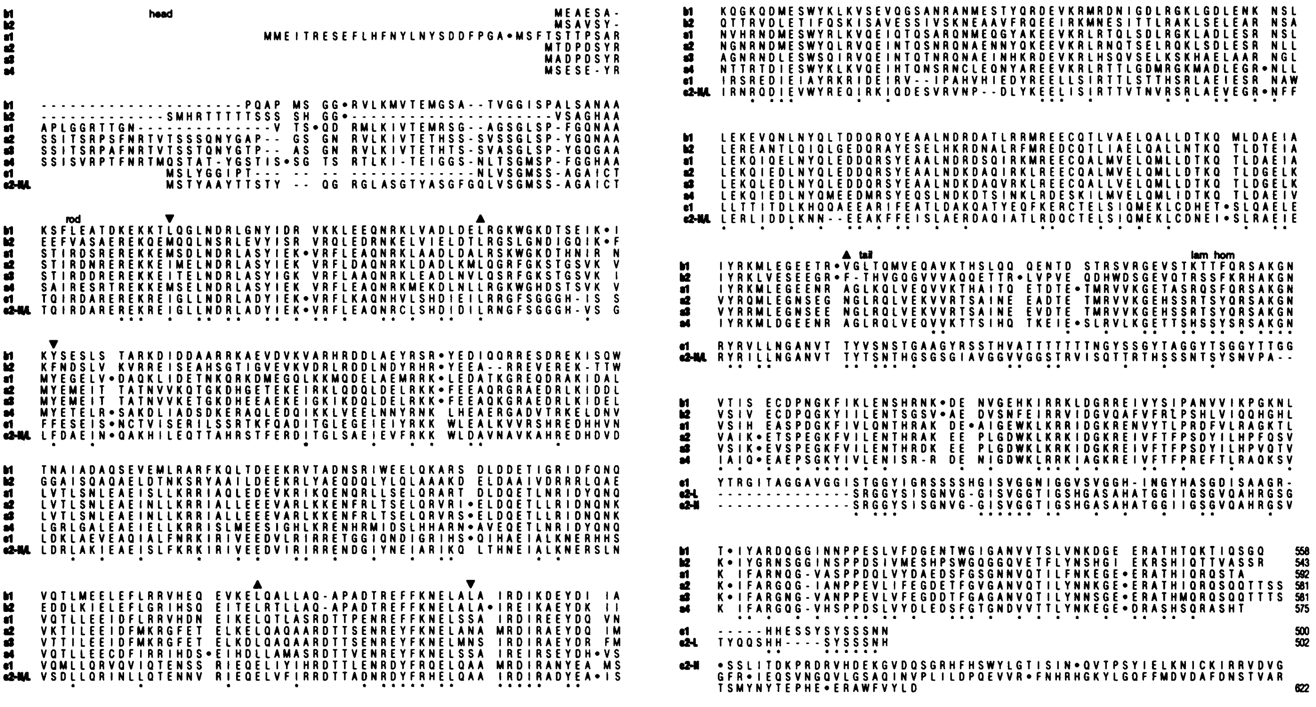

Fig. 5. Alignment of *C.elegans* IF protein sequences and intron positions of the corresponding genes. Gaps introduced for optimization of the alignments are represented by dashes. The three structural domains (head, rod, tail) are indicated. The starts and ends of the α-helical subdomains (coils 1a, 1b and 2) of the rods are marked by down- and up-pointing arrowheads, respectively. All eight head and rod sequences are grouped into a single alignment. The tail domains are displayed as two groups based on the presence (a and b proteins) or absence (c proteins) of a region with substantial sequence homology with nuclear lamins. '*Lam hom*' marks the onset of this region (~120 residues), which entends to the C-terminal end of the IF a and b proteins (see Weber *et al.*, 1989). These regions share 23–34% sequence identity with the corresponding tail sequences of mouse lamin B2 (Zewe *et al.*, 1991), *Xenopus* lamin B3 (Döring and Stick, 1990) or *Drosophila* lamin $Dm_0$ (Osman *et al.*, 1990). Asterisks below the bottom lines mark identical or homologous residues that occur in all aligned sequences. The following groups of amino acids have been used for residue similarity: AILMV/FYW/DE/HKR/ST/NQ. Note the high homology along the entire coil 1a regions and the C-terminal halves of the linkers between the coil 1b and coil 2 subdomains. While the N-terminal portions of the coil 1a subdomains follow the consensus-type IF sequence, several strong variations on the canonical YRKLLEGEESR sequence motif at the C-terminal end of the rod domain are seen. Apart from a few very short deletions (see Results) all eight rod domains are essentially of identical length, as are the tail domains of the IF a and b proteins. The total number of residues for each IF polypeptide is given at the end of its sequence. Note that the last 15 residues, starting with Gly488, of the tail domain of $c_2$-L are substituted in $c_2$-H by an extension of 135 residues which begin with Val488 (see Results). The $a_1$ sequence shown corresponds to the largest possible head domain, $a_1$-H. The $b_2$ sequence shown represents $b_2$-H. The first 17 residues of the tail domain are eliminated in $b_2$-L. Dots within the protein sequences mark intron positions in the corresponding genes. Dots following an amino acid residue do not distinguish whether the intron is located after the first (phase 1), second (phase 2) or third (phase 3) nucleotide of the corresponding codon. For documentation of intron phases see Figure 10. The DNA sequences are available from EMBL/GenBank database under accession numbers X70830–X70836, X78553 and X78554.

central α-helical rod domain involved in coiled coil dimer formation and the flanking non-helical head and tail domains (Figure 5). The rod domains show the characteristic segmentation into coils 1a, 1b and 2 which are separated by short non-helical spacers. Since the tail domains lack the hallmark motifs of nuclear lamins, i.e. the nuclear localization signal and the C-terminal CaaX box, all proteins analyzed represent bona fide cytoplasmic IF proteins. Whereas all eight rod domains are of similar length (~360 residues) the terminal regions differ in size. Head domains range from 25 to 78 residues and the tail domains vary between 117 and 223 residues. The complete polypeptides comprise between 500 and 622 residues corresponding to calculated molecular masses of 55 868 ($c_1$) to 70 101 ($c_2$-H). Calculated isoelectric points are between 5.46 ($b_1$, $b_2$-L) and 7.02 ($c_1$) (see Table I).

All *C.elegans* IF proteins have the extended coil 1b subdomain found in nuclear lamins (Fisher *et al.*, 1986; McKeon *et al.*, 1986) and the previously characterized cytoplasmic IF proteins of invertebrates (Weber *et al.*, 1988,

Table I. Physical parameters of *C.elegans* IF proteins

| Cel IF protein | Residues | Mol. wt | pI |
|---|---|---|---|
| $b_1$ | 558 | 63 700 | 5.46 |
| $b_2$-H | 543 | 61 695 | 5.55 |
| $b_2$-L | 526 | 59 856 | 5.46 |
| $a_1$-H | 592 | 68 547 | 5.93 |
| $a_1$-M | 575 | 66 530 | 6.30 |
| $a_1$-L | 567 | 65 538 | 6.48 |
| $a_2$ | 581 | 67 094 | 6.00 |
| $a_3$ | 581 | 66 562 | 5.90 |
| $a_4$ | 575 | 66 359 | 6.64 |
| $c_1$ | 500 | 55 868 | 7.02 |
| $c_2$-H | 622 | 70 101 | 6.74 |
| $c_2$-L | 502 | 56 207 | 6.84 |

All *C.elegans* IF proteins characterized in this study are listed by their names, residue numbers, calculated molecular weights and theoretical isoelectric points. Designations H, M and L refer to the isoforms generated by differential RNA splicing pathways of the corresponding primary transcripts (see Results).

ROD SEQUENCE IDENTITY

TAIL SEQUENCE IDENTITY

| | b1 | b2 | a1 | a2 | a3 | a4 | c1 | c2 |
|---|---|---|---|---|---|---|---|---|
| b1 | | 49 | 56 | 46 | 44 | 42 | 21 | 25 |
| b2 | 38 | | 40 | 34 | 33 | 33 | 21 | 25 |
| a1 | 54 | 38 | | 68 | 63 | 64 | 33 | 39 |
| a2 | 52 | 34 | 64 | | 86 | 57 | 34 | 37 |
| a3 | 52 | 34 | 66 | 90 | | 53 | 32 | 34 |
| a4 | 47 | 35 | 64 | 54 | 56 | | 32 | 35 |
| c1 | NA | NA | NA | NA | NA | NA | | 64 |
| c2 | NA | NA | NA | NA | NA | NA | 32 | |

**Fig. 6.** Sequence comparison of rod and tail domains of *C.elegans* IF proteins. Sequence identity values (%) are summarized for the various rod and tail domains. Because of their entirely different sequences the tail domains of $c_1$ and $c_2$ have not been considered for comparison (NA: not applicable). The two members of the b subfamily display a strong sequence drift. The grouping of $b_2$ next to $b_1$ into the same subfamily is largely based on the high homology in exon/intron organization of the corresponding genes (see Figures 5 and 10). The rod domains of $c_1$ and $c_2$ show a closer relation to the IF a proteins than to the IF b proteins. The overall sequence homologies for all eight rod domains and the six lamin-like tail domains of IF a/b proteins are 29 and 40%, respectively.

1989; Szaro *et al.*, 1991). While this extension is usually precisely six heptads or 42 residues, the *C.elegans* protein $b_2$ has only 39 residues due to two short internal deletions (Figure 5). Unusual deletions can also occur in the coil 2 subdomain. *C.elegans* proteins $c_1$ and $c_2$ both have a deletion of two residues and a flanking proline residue in the N-terminal half of coil 2. This is the same region where filensin, the vertebrate IF protein of lens fiber cells (Gounari *et al.*, 1993; Remington, 1993), and a nuclear lamin of *C.elegans* (Riemer *et al.*, 1993) display larger deletions of 29 and 14 residues respectively. A further gap of two residues occurs in the *C.elegans* $c_2$ protein in the central part of the coil 2 subdomain (Figure 5). Interestingly, this site coincides with the position of the heptad reversal or stutter in the coil 2 subdomain (see Steinert and Roop, 1988).

Based on sequence similarities along the rod and tail domains the *C.elegans* IF proteins can be divided into three distinct subfamilies, referred to as CelIF a, b and c. This subdivision is supported by the general exon/intron organizations of the corresponding genes (see below) and the existence of group-specific intron positions which are unique to and shared by all members of a given subfamily. Indeed, the latter criteria group protein $b_2$ next to $b_1$ into the same subfamily while by sequence homology alone $b_2$ could be considered a special type, separate from $b_1$ and the a subfamily. Figure 6 summarizes the sequence identity values for the various *C.elegans* IF proteins. The overall sequence homology along the rod domains of the *C.elegans* proteins is particularly high in the entire coil 1a subdomain, the N-terminal region of coil 2 and the second half of the linker between coil 1b and 2. An increase in homology is again seen at the C-terminal end of the rod. The lowest homology is observed along the coil 1b segments (Figure 5).

The non-neuronal IF proteins of gastropods and the two muscle IF proteins of *Ascaris* display an extended region in their tail domains which shows ~30% sequence identity with the corresponding regions of nuclear lamins (Weber *et al.*, 1988, 1989; Dodemont *et al.*, 1990; Riemer *et al.*, 1991). This relation is also seen in one of the three neurofilament proteins of the squid (Szaro *et al.*, 1991; Way *et al.*, 1992). This sequence homology with lamins along the tail domain is restricted to the members of the a and b subfamilies of *C.elegans* IF proteins. The proteins $c_1$ and $c_2$ have totally unrelated tail domains (Figure 5; see below). Interestingly, their rod domains also diverge considerably from those of the a and b subfamilies (Figure 6).

Additional sequence diversity is generated for the head domain of $a_1$ and the tail domains of $b_2$ and $c_2$ by alternative splicing and processing of the corresponding RNA transcripts (see below). At least three distinct variations occur for the head domain of $a_1$ as $a_1$-H, M and L. The tail domain of $b_2$ appears either at full size in $b_2$-H (Figure 5) or as a shortened form in $b_2$-L due to the elimination of 17 residues by exon skipping. The tail domain of $c_2$-L can be extended by a unique sequence of 135 residues giving rise to $c_2$-H (Figure 5).

The lamin unrelated tail domains of $c_1$ and $c_2$-L show a high content of glycine and serine residues (~43%). Negatively charged residues are either completely absent ($c_2$-L) or very rare (two residues in the $c_1$ tail domain). Although most glycine and serine residues occur as part of a sequence motif repeated 12 or 13 times, they are not organized into arrays of consecutive residues flanked by hydrophobic amino acids as in various vertebrate keratins (Marchuk *et al.*, 1984; Johnson *et al.*, 1985; Krieg *et al.*, 1985; Tyner *et al.*, 1985; Krauss and Franke, 1990). Instead the motif centers around the sequence (G)GXXX with one or two glycines followed by three amino acids with hydrophobic or uncharged polar side chains (A, G, I, N, S, T, V or Y). The extension of the tail domain in $c_2$-H represents unique sequences with no counterparts in the data banks; it shows no obvious motifs or repeats.

### *Trans*-splicing of the IF RNA transcripts

All IF-specific pre-mRNAs, except those arising from the $c_2$ gene, undergo *trans*-splicing to generate the 5′-ends of the mature mRNAs. This could be directly inferred from the 5′-sequences of the cDNAs encoding $a_4$, $b_1$, $b_2$ and $c_1$, all of which contained remnants (the last six to 20 bp) of the 22 nt *trans*-spliced SL1 leader sequence (Krause and Hirsh, 1987). For all other mRNAs except the $c_2$ mRNAs this was demonstrated by the successful PCR rescue of their very 5′-ends employing sense primers that were identical to the SL1 leader sequence. Accordingly, in all genes except for $c_2$, acceptor sites selected for *trans*-splicing are located at the appropriate positions (see Figure 3). All but one site occur at short distances of 2−16 nt upstream to the designated translational initiation codons, which is well within the range reported for other *C.elegans* genes (Bektesh *et al.*, 1988; Agabian, 1990).

The $a_1$ pre-mRNAs show *trans*-splicing at alternative splice acceptor sites whose selection is probably under control of two distinct promoters (Figure 7A and B). Similar characteristics have been described for the *unc-5* transcripts of *C.elegans* (Leung-Hagesteijn *et al.*, 1992). *Trans*-splicing to either of two acceptor sites located immediately upstream of the first ATG codons in exon 1 or exon 2, is accompanied

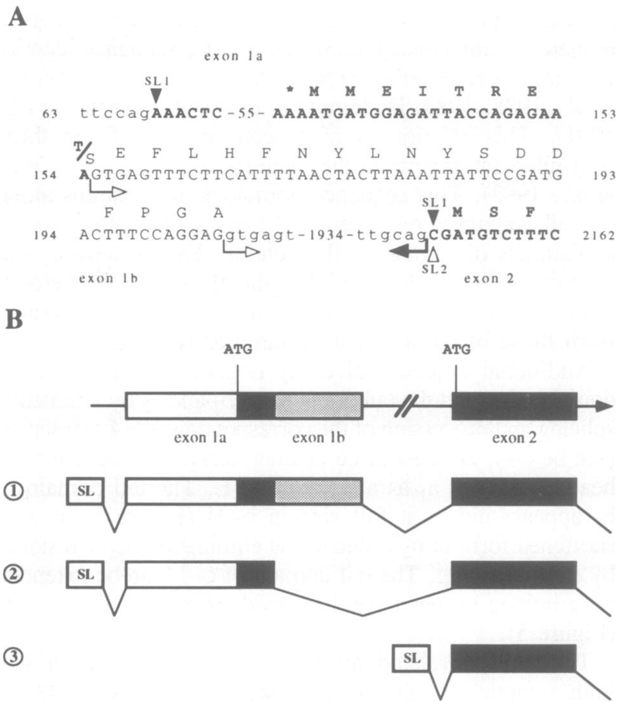
A
exon 1a
SL1
* M M E I T R E
63 ttccagAAACTC-55- AAAATGATGGAGATTACCAGAGAA 153
T/S E F L H F N Y L N Y S D D
154 AGTGAGTTTCTTCATTTTAACTACTTAAATTATTCCGATG 193
F P G A
SL1
M S F
194 ACTTTCCAGGAGgtgagt-1934-ttgcagCGATGTCTTTC 2162
SL2
exon 1b
exon 2
B
ATG
ATG
exon 1a
exon 1b
exon 2
①
SL
②
SL
③
SL

**Fig. 7.** Alternative *cis*- and *trans*-splicing pathways of CelIF $a_1$ RNA transcripts. (**A**) The 5′-sequence of the $a_1$ gene comprising exon 1, intron 1 and the 5′-end of exon 2. Upper case letters represent exon sequences; bold and normal face letters define compulsory and optional exon sequences, respectively. Lower case letters specify intron or (upstream to *trans*-splice acceptor sites) 'outron' (Spieth *et al.*, 1993) sequences. The *cis*-splicing intron 1 is marked at its fixed 3′-end by a thick arrow. Each of the two alternative 5′-ends is indicated by thin arrows. Filled and open arrowheads denote the acceptor sites for the *trans*-spliced leader sequences SL1 and SL2 (Krause and Hirsh, 1987; Huang and Hirsh, 1989), respectively. The asterisk marks the 5′-end of the largest $a_1$ cDNA isolated from the library. Another $a_1$ cDNA starts with the SL1 sequence followed by exon 2. (**B**) Diagrammatic presentation of splicing routes. Exons are depicted as rectangles and drawn to scale. Open boxes indicate 5′-non-coding regions. Black and half-tone boxes represent compulsory and optional coding sequences, respectively. The three pathways, 1–3, give rise to distinct $a_1$ mRNAs encoding large (H), medium (M) and small (L) head domains, respectively. While all RNA transcripts undergo *trans*-splicing to SL1, the small transcript generated by pathway 3 can also splice in *trans* to SL2.

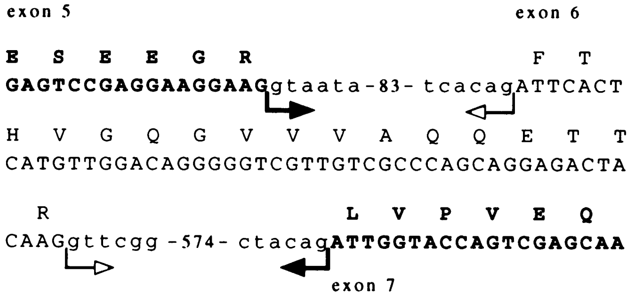
exon 5
exon 6
E S E E G R
F T
GAGTCCGAGGAAGGAAGgtaata-83- tcacagATTCACT
H V G Q G V V V A Q Q E T T
CATGTTGGACAGGGGGTCGTTGTCGCCCAGCAGGAGACTA
R
L V P V E Q
CAAGgttcgg -574- ctacagATTGGTACCAGTCGAGCAA
exon 7

**Fig. 8.** Alternative *cis*-splicing of CelIF $b_2$ RNA transcripts: exon skipping. The $b_2$ gene region around the boundary of rod- and tail-specific sequences is shown. Designations of the different letter types are as in Figure 7. Thick and thin arrows denote compulsory and optional intron border sequences, respectively. Recruitment of the optional acceptor and donor splice sites from the large intron results in the incorporation of the mini exon 6 encoding the first 17 amino acids of the tail domain. Alternatively, splicing of exon 5 to exon 7 leads to the elimination of this exon (exon skipping). The resulting mature $b_2$-H and $b_2$-L mRNAs differ only with respect to the presence or absence of exon 6.

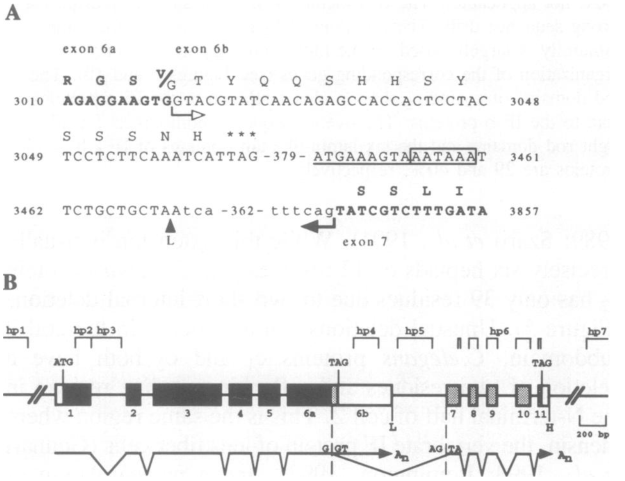
A
exon 6a
exon 6b
R G S V/G T Y Q Q S H H S Y
3010 AGAGGAAGTGGTACGTATCAACAGAGCCACCACTCCTAC 3048
S S S N H ***
3049 TCCTCTTCAAATCATTAG -379- ATGAAAGTAAATAAAT 3461
S S L I
3462 TCTGCTGCTAAtca -362- tttcagTATCGTCTTTGATA 3857
L
exon 7
B
hp1
hp2 hp3
hp4
hp5
hp6
hp7
ATG
TAG
TAG
1
2
3
4
5
6a
6b
7
8
9
10
11
L
H
200 bp
G|GT
AG|TA
An
An

**Fig. 9.** Alternative processing of CelIF $c_2$ RNA transcripts using different polyadenylation signals. (A) Only the region of the $c_2$ gene which shows the divergence of the tail domain encoding sequences of the $c_2$-L and $c_2$-H RNAs is shown. Designations of the different letter types and arrows are as in Figures 7 and 8. The $c_2$-L transcripts recruit from intron 7 either of three consecutive polyadenylation signals, ATGAAA, AGTAAA and AATAAA (underlined), of which the latter signal (framed) displays the canonical sequence (Proudfoot and Brownlee, 1976). The arrowhead marks the poly(A) addition site of $c_2$-L RNA. (**B**) Diagrammatic presentation of splicing pathways. Exons (rectangles) and introns (lines) are drawn to scale. Coding sequences common to both RNA transcripts are depicted in black whereas those unique to either transcript are given in half-tone. Open boxes specify 5′- and 3′- untranslated sequences. The positions of the polyadenylation sequences utilized by the $c_2$-L and $c_2$-H RNA transcripts are indicated. Hybridization probes (hp 1–7) used in Northern analysis (see Results) are delineated at the top.

by alternative *cis*-splicing. The concomitant splicing events generate at least four different $a_1$ RNA transcripts. Two of these were directly isolated from the library as the corresponding cDNAs whereas the other two were identified by RT-PCR on total mRNA. The two largest RNA transcripts both *trans*-splice to SL1. Whereas one species ($a_1$-H RNA) contains all of exon 1, the other ($a_1$-M RNA) selects a new 5′-*cis*-splice site for intron 1, shifted upstream of the $a_1$-H specific position, resulting in the elimination of the 3′-part of exon 1. The two shorter RNA transcripts ($a_1$-L RNAs) starting with exon 2 are identical except that they splice *in trans* to either SL1 or SL2 RNA (Huang and Hirsh, 1989). The *trans*-spliced RNA transcripts give rise to three different head domains for the $a_1$ protein. It is highly unlikely that the third ATG codon of $a_1$-H mRNA will be eligible for translational initiation (Kozak, 1991). Thus *trans*-splicing, which recruits this codon as the first for $a_1$-L mRNA, provides a mechanism to bypass the otherwise largely preferred first (or second) ATG codon of exon 1 in favor of a downstream internal translational start signal.

### *Exon skipping and alternative RNA processing involving the tail domains of CelIF $b_2$ and $c_2$*

The two distinct $b_2$ cDNAs correspond to mRNAs that differ with respect to the presence ($b_2$-H) or absence ($b_2$-L) of a small exon defining the first 17 amino acids of the tail domain. Figure 8 shows the alternative RNA splicing

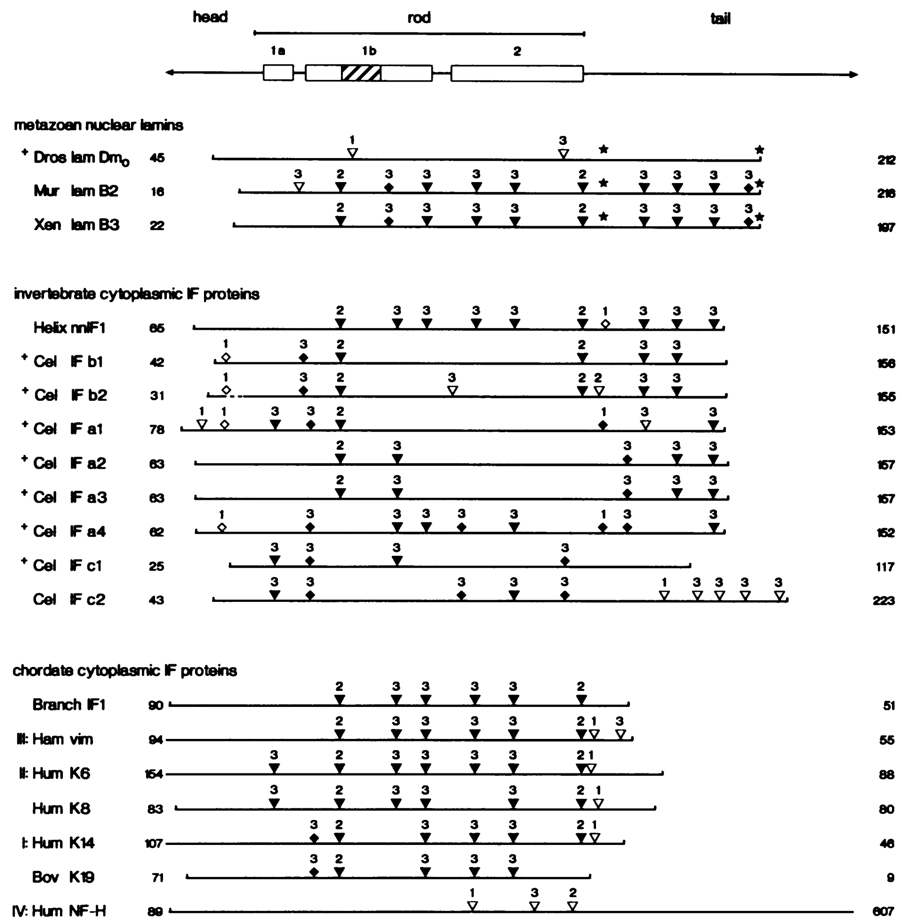

**Fig. 10.** Summary of intron positions in the lamin/cytoplasmic IF multigene family. The tripartate structural organization of the proteins is indicated at the top. Residue numbers for the variable head and tail domains are given at the sides. The hatched box in coil 1b marks the additional 42 residues (six heptads) unique to all nuclear lamins and the currently known cytoplasmic IF proteins of invertebrates. Asterisks in the lamin tail domains mark the positions of the nuclear localization signals and CaaX boxes. Plus signs at the names of all but one of the *C.elegans* IF genes indicate the presence of *trans*-splice acceptor sites, e.g. 3′-boundaries of introns ('outrons', see Spieth *et al.*, 1993) preceding the translational initiation codons. The plus sign by the *Drosophila* lamin gene name refers to the *cis*-splicing intron in the 5′-untranslated sequence. All other intron positions are given as triangle or diamond symbols relative to the protein coding sequences (see domain sketch at the top and Figure 5). Numbers above the symbols indicate the phase of the intron. Phases 1, 2 and 3 refer to interruptions after the first, second and third nucleotide of a codon, respectively. Solid symbols mark intron positions which are strictly conserved with respect to both the amino acid position and the codon phase. Solid triangles denote those introns which are shared by genes from at least two of the three gene families (nuclear lamin genes, cytoplasmic IF genes of invertebrates and chordates respectively). Solid diamonds mark identical intron positions found only in different genes from a single gene family. Open triangles refer to unique intron positions not present in any other gene. Open diamonds indicate intron positions which, although corresponding in location and phase to those in other genes from the same family, reside in slightly different sequence contexts. The latter intron type is represented by the introns in the head domains of the *C.elegans* $b_1$, $b_2$, $a_1$ and $a_4$ genes, which may have arisen from a single common position by intron sliding mechanisms. Another example is the position of the seventh intron of the *Helix* gene versus the two identical intron positions (solid diamonds) of the *C.elegans* $a_1$ and $a_4$ genes. Note the remarkable drift of intron patterns among the *C.elegans* IF genes. Each gene shares only one to four intron positions with the *Helix* gene. Nevertheless, together the intron positions of the eight *C.elegans* IF genes cover all but the fourth intron of the *Helix* gene. Note also the introns in the coil 1a subdomain of the $a_1$, $c_1$ and $c_2$ genes. They occur in precisely the same location as the first intron of vertebrate type II keratin genes. Intron positions are given for the genes encoding *Drosophila* lamin $Dm_0$ (Osman *et al.*, 1990), mouse lamin B2 (Zewe *et al.*, 1991), *Xenopus* lamin B3 (Döring and Stick, 1990), *Helix* non-neuronal IF proteins (Dodemont *et al.*, 1990), *C.elegans* IF proteins $b_1$, $b_2$, $a_1$, $a_2$, $a_3$, $a_4$, $c_1$ and $c_2$ (this study), *Branchiostoma* IF protein (Riemer *et al.*, 1992), hamster vimentin (Quax *et al.*, 1983), human keratin 6 (Tyner *et al.*, 1985), human keratin 8 (Krauss and Franke, 1990), human keratin 14 (Marchuk *et al.*, 1984), bovine keratin 19 (Bader *et al.*, 1986) and human high molecular weight neurofilament protein (Lees *et al.*, 1988). The figure lists only major prototypes for vertebrate type I to III genes; for other genes see GenBank. For other type IV neurofilament genes, all of which share the second and third intron of the NF-H gene, see references cited by Lees *et al.* (1988). The nestin gene also shows a type IV intron pattern (Dahlstrand *et al.*, 1992).

pathways that lead to the inclusion or elimination of the mini exon. This exon is part of a 732 bp sequence which can be removed as a single intron.

In contrast to the $b_2$ mRNAs, the two different $c_2$ mRNAs show tail domain encoding sequences which are identical for the N-terminal regions but diverge completely over the C-terminal regions (Figure 5). The 3′-part of the single large exon specifying the tail domain of $c_2$-L is exchanged in $c_2$-H mRNA by a unique extended sequence encoded by five additional short exons (Figures 9 and 10). Both mRNAs with identical 5′-ends but different 3′-termini are generated from the single $c_2$ gene by alternative RNA processing resulting from differential utilization of two polyadenylation signals spaced ~ 1.1 kb apart (Figure 9B). Similar RNA processing pathways have been described for the gene encoding non-neuronal IF proteins of *H.aspersa* (Dodemont *et al.*, 1990). The $c_2$-L mRNA is spliced from the shorter transcript which selects the polyadenylation signal located within intron 6. Exon 6a is thus committed to extend with the 5'-part of intron 6 as exon 6b starting with the glycine codon as onset of a short (15 amino acid residues) tail domain extension. The accessibility of the 3'-splice acceptor site of intron 6 by transcriptional termination beyond the second polyadenylation signal sequence gives rise to a splicing pathway which produces $c_2$-H mRNA. Here exon 6a joins to exon 7 which splices to the four additional exons 8–11. Together exons 7–11 encode a unique tail extension of 135 amino acid residues, of which the first is a valine rather than a glycine residue (see above). As judged from Northern analysis (not shown) using $c_2$-L (probe 4) and $c_2$-H mRNA (probe 6) specific hybridization probes (Figure 9B), the fully spliced mRNAs are present at similar levels in poly(A)$^+$ RNA from mixed-stage nematode populations. Probe 4 detected in addition several minor RNA transcripts of ~4 kb (see Figure 4) which also hybridized to all upstream located probes tested (probes 1–3). These were not detected by probes derived from regions downstream of the $c_2$-L mRNA poly(A) addition site (probes 5–7). The large RNA species may therefore represent unspliced precursor RNAs for the $c_2$-L mRNA. Alternatively they may arise from a separate transcription unit overlapping with the 5'-part of the $c_2$ gene in either the same or the opposite direction. Given the high gene density of *C.elegans* such a constellation cannot currently be excluded (see Sulston *et al.*, 1992).

### *Highly divergent exon/intron organizations of the C.elegans IF genes*

The structural organizations of the IF genes show a very strong drift both in number and in position of the introns. Not a single intron position is shared by all eight genes. While individual genes contain between four ($c_1$) and 10 ($c_2$) introns, the multigene family displays a total of 55 introns that occur in as many as 28 distinct positions (Figure 10). Most introns interrupt sequences encoding the rod and tail domains. Genes $b_1$, $b_2$, $a_1$ and $a_4$ are the first cytoplasmic IF genes described which show one or even two introns interrupting the head domain specific sequences. Of these five introns, four occur in similar but not identical positions and have the same phase (Figures 5 and 10). These introns and an additional eight are unique introns in the current IF gene catalog (Figure 10). All other intron positions appear in at least two genes.

Even the different members of the same gene subfamily show rather divergent intron patterns, with the noticeable exception of genes $a_2$ and $a_3$ which also share by far the highest sequence identity (see Figures 3 and 6). Nevertheless the exon/intron organizations provide two important features supporting the subdivision originally made by sequence criteria alone (see above). First, all members of a subfamily share intron positions, which do not occur in other subfamilies. Second, the tail domain homology versus the nuclear lamins present in a and b genes is reflected by intron patterns which together cover all but one intron position in the tail domain of vertebrate B-type lamin genes, e.g. the *Xenopus* lamin B3 and mouse lamin B2 genes (Döring and Stick, 1990; Zewe *et al.*, 1991). In contrast the complete lack of lamin homology in the tail domains of the $c_1/c_2$ genes parallels the totally divergent structural organizations for these parts of the genes. Introns are either completely absent ($c_1$ gene) or occur in unique positions exclusively related to the extension of the tail domain ($c_2$ gene).

Individual *C.elegans* IF genes share only a limited number of identically placed introns with the *H.aspersa* non-neuronal IF gene (Dodemont *et al.*, 1990). While five genes ($a_2$, $a_3$, $a_4$, $b_1$ and $b_2$) each have four intron positions in common with the *Helix* gene, the latter shares two introns with the $a_1$ gene and only a single intron location with either the $c_1$ or $c_2$ gene. However, together the eight *C.elegans* IF genes have exact counterparts for eight of the 10 intron positions of the *Helix* IF gene (Figure 10). The position of intron 7 of the gastropod gene is similar but not identical to corresponding intron locations in genes $a_1$ and $a_4$ which occur in a region of noticeable length variability. Since all these introns share phase 1 versus the reading frame they are probably homologous introns, particularly when intron sliding mechanisms are considered. Thus nine of the 10 introns of the *Helix* IF gene are covered by the *C.elegans* IF genes in spite of their striking diversity in intron patterns. Only intron 4 of the gastropod gene has no direct counterpart in the *C.elegans* IF gene family.

Conversely, the *C.elegans* IF genes display 19 intron positions, none of which occur in the *Helix* gene. Only one of these is present in other metazoan IF genes (see below), but all other intron positions are entirely novel. Thirteen of these are single introns, even among the *C.elegans* IF genes themselves. Some of the introns are of particular interest. These include the unique introns interrupting the coding sequences of the head domains in four genes (see above) and the intron located in the highly conserved coil 1a sequences of genes $a_1$, $c_1$ and $c_2$. It occupies precisely the same position as the first intron of vertebrate type II keratin genes (Figure 10; see Krauss and Franke, 1990, for a review). The current catalog of *C.elegans* IF genes lacks, however, an intron which corresponds precisely to the first intron of vertebrate type I keratin genes (see Bader *et al.*, 1986 for a review). An intron common to genes $a_1$, $a_4$, $c_1$ and $c_2$ resides 15 bp 5' to the location of this particular intron. Finally, the unique sixth intron of the $b_2$ gene has the same phase as the preceding intron, which coincides with the onset of the tail domain. This allows skipping of the interjacent mini exon by alternative RNA splicing (see above).

## Discussion

### *Diversity of C.elegans IF sequences*

The cytoplasmic intermediate filament proteins of the nematode *Caenorhabditis elegans* are encoded by a multigene family of a minimum of eight genes which comprise three distinct (a, b and c) subfamilies. The complexity of IF protein composition is increased to at least 12 distinct species (Table I) by an array of mechanisms acting at the transcriptional and post-transcriptional levels. These include differential promoter use leading to alternative RNA *trans*-splicing pathways, as well as alternative RNA *cis*-splicing and polyadenylation events, all of which account for extra variations on the non-helical end domains. Different variants of single rod domains have not been found. Three distinct head domains have been identified for the $a_1$ protein ($a_1$-H, M and L) and two different tail domains for the $b_2$ ($b_2$-H and L) and $c_2$ ($c_2$-H and L) proteins. The actual IF diversity may be higher than anticipated from the currently isolated genes and cDNAs. First, there may be more than eight IF genes. Thus far, not all fragments detected by Southern blot hybridization of genomic DNA digests with the various *Ascaris* and *C.elegans* probes (see Figure 1) have been analyzed systematically. Some may represent spurious hybridization signals or even pseudogenes which have also been found for other multigene families of *C.elegans* (Ward *et al.*, 1988). On the other hand, some IF-specific sequences may have escaped detection. This is illustrated by the $b_2$ gene, which was only identified by the corresponding EST sequences (Waterston *et al.*, 1992) and subsequently cloned in this study. Second, additional RNA splice variants with still other insertion and/or deletion features related to head, tail or even rod domains may exist. Proof of such variants will require the isolation of the corresponding cDNAs either directly from libraries or by recovery as RT-PCR products depending on their relative abundance in total mRNA. Typical examples in the present study are the $b_2$-L and $a_1$-M mRNAs. The expression level of $b_2$-L mRNA was sufficiently high for both us and Waterston *et al.* (1992) to be able to isolate the corresponding cDNA directly. In contrast, the rare $a_1$-M RNA variant was identified by chance during RT-PCR analysis of the *trans*-spliced $a_1$-H and $a_1$-L transcripts. Thus, exhaustive RT-PCR testing will be required to reveal the full range of IF sequence diversity in *C.elegans*.

The large IF sequence diversity of *C.elegans* is astounding for an animal with a simple body plan (see Wood, 1988). Northern analysis shows that *C.elegans* IF mRNAs are expressed at very different levels in mixed-stage populations. Therefore some mRNAs may be developmentally regulated. They could either be embryo, larval or adult specific and/or confined to certain cell types of a given developmental stage. Previous immunological studies with several monoclonal antibodies including IFA revealed the presence of putative IFs in distinct tissues and cell types including the body wall, the uterus, the marginal cells of pharyngeal epithelium and the excretory cell (Bartnik *et al.*, 1986; Francis and Waterston, 1991). These studies can now be extended with the elucidation of temporal and spatial expression patterns of individual IF genes. Furthermore, the collection of IF cDNAs should allow IF sequences to be related to the specific antibodies known (Francis and Waterston, 1991).

### Evolution of the IF multigene family

The eight genes encoding the *C.elegans* IF proteins show strikingly distinct intron patterns and do not have a single common intron position. These genes illustrate nicely the pitfalls in the interpretation of intron pattern homology when only one gene of a multigene family is compared with another gene from a different phylum. Homologies in gene organizations can be highly over- or highly underestimated. While genes $c_1$ and $c_2$ of *C.elegans* share only a single intron with the *H.aspersa* non-neuronal IF gene, five genes ($a_2$, $a_3$, $a_4$, $b_1$ and $b_2$) display different subsets of four identically placed introns. Even more striking is the finding that nine of the 10 introns of the gastropod gene have a counterpart in one or the other member of the *C.elegans* IF gene family. Only intron 4 of the gastropod gene lacks a homolog in the current catalog of *C.elegans* IF genes (Figure 10). Nevertheless this intron position has not been lost during the evolution of the nematodes since it was found in an IF gene of the large nematode *A.lumbricoides* (our unpublished results). The current results are compatible with the assumption that the common progenitor of molluscs and nematodes displayed in its IF gene all 10 intron positions which are present in the *Helix* gene. During the multiple gene duplication events leading to the nematode IF gene family these introns were lost to different degrees in the various progeny genes. Thus massive but individually restricted intron losses accompanied the evolution of the *C.elegans* IF gene family.

The remaining 19 intron positions in the *C.elegans* catalog represent either old introns, which were lost along the lineage towards the molluscs, or new introns acquired specifically during the evolution of the branch leading to the nematodes and their multiple IF genes. Alternatively only some of these introns may be old and expected to be present in the progenitor of molluscs and nematodes, while others were gained on the nematode branch. Currently a few of these 19 introns seem of particular interest. The finding that three *C.elegans* genes contain one or even two introns in the head domain may offer an explanation for a usual feature of IF gene organization. With the exception of the *C.elegans* genes $b_1$, $b_2$ and $a_4$, all other IF genes described so far show unusually large sizes of the first exon (Figure 10). Thus we speculate that the archetypal IF gene may have had homologous intron(s) in the head domain and that these were lost on other evolutionary branches but retained in some *C.elegans* genes. The last intron in the rod domains of genes $c_1$ and $c_2$ lies six nucleotides past an intron of the *Drosophila* lamin $Dm_0$ gene (Figure 10), which previously had no counterpart in the developing intron catalog of IF genes (Osman *et al.*, 1990). The intron present in a highly conserved region of the coil 1a domain is found in three *C.elegans* genes ($a_1$, $c_1$ and $c_2$) and in all vertebrate type II keratin genes. Given the arguments of Dibb and Newman (1989) this intron could have been independently acquired on different metazoan radiations due to the presence of a proto-splice site in the coding sequence. Alternatively this intron could also be an old intron, which was kept in some nematode genes but lost in others and in the gastropod gene. During the evolution of the predecessors of vertebrate type I to IV genes it was lost in most predecessors except that of the keratin type II genes. Decisions about intron gain and loss models for this and for some other introns must be postponed until more invertebrate IF genes are known.

Invertebrate IF proteins show, like all nuclear lamins, 42 extra residues in the coil 1b domain and generally display a lamin-like homology segment in the tail domain. Thus it is thought that cytoplasmic IF proteins arose in eukaryotic evolution from a mutated lamin gene due to the loss of the nuclear localization signal and the CaaX motif (Dodemont *et al.*, 1990; Döring and Stick, 1990; Weber *et al.*, 1988, 1989). The characterization of the eight *C.elegans* IF genes indicates a two-step mechanism of domain evolution. All *C.elegans* IF proteins have retained the long coil 1b version (for a very minor change in gene $b_2$ see Figure 5 and Results), but only six genes give rise to proteins with a lamin-like tail domain. The precursors of genes $c_1$ and $c_2$ have acquired, possibly by exon shuffling, entirely distinct tail domains. Thus, rod and tail domains seem under different evolutionary constraints in lower metazoa. This asynchronous evolution of the two domains is also evident in the sequences of three neurofilament proteins of the squid

*Loligo pealei*. They arise from a single gene and share the same long coil 1b domain (Szaro *et al.*, 1991; Way *et al.*, 1992). Inspection of the sequences shows, although not stated by the authors, that the NF70 protein contains the lamin-like tail domain while the NF60 and NF200 proteins have gained novel tail domains. Thus, lamin homology of the tail domain can be lost even by cytoplasmic IF genes that still retain the long coil 1b domain. We expect that the shortening of the coil 1b domain occurred at some, as yet unknown, stage during the evolution of the predecessors of chordate type I to IV IF genes (Szaro *et al.*, 1991; Riemer *et al.*, 1992).

## Materials and methods

### *Extraction and analysis of genomic DNA and mRNA*

Mixed-stage populations of *C.elegans* (wild-type strain N2 Bristol) were used throughout this study. Growth, maintenance and harvesting were as described (Sulston and Hodgkin, 1988). Nematodes were either shock-frozen in liquid nitrogen and stored at −80°C or immediately used for preparation of nucleic acids. Genomic DNA was isolated essentially as described (Sulston and Hodgkin, 1988). Preparation of total cellular RNA was as follows. Three grams of frozen nematodes were ground in a mortar on liquid nitrogen. The powdered material was homogenized at 22°C with a Polytron blender set at moderate speed in 30 ml of freshly prepared buffer consisting of 50% (v/v) Tris-base saturated phenol in 50 mM Tris−HCl, pH 8.0, at 22°C, 20 mM CDTA (1,2-cyclohexylenedinitrilotetraacetic acid monohydrate, Sigma) and 1% Na-sarcosyl. Subsequent purification and recovery of nucleic acids followed standard procedures (Sambrook *et al.*, 1989). Total RNA (16 mg yield) was freed of contaminating DNA and polysaccharides by two extractions with ice-cold 3 M sodium acetate, pH 6.0 (Palmiter, 1974). Poly(A)$^+$ RNA (300 μg yield) was isolated by two cycles of affinity chromatography on oligo(dT)−cellulose (Dodemont *et al.*, 1990). All subsequent manipulations of the nucleic acids isolated were performed using established techniques (Sambrook *et al.*, 1989). Hybridization at high or low stringency of Southern and Northern blots was as detailed below.

### *Isolation and characterization of genomic and cDNA clones*

The first genomic clones were isolated from a genomic library of partial *Sau*3A-digested *C.elegans* DNA in the phage vector λ2001 (Karn *et al.*, 1984; Coulson *et al.*, 1986), which was kindly provided by Dr John Sulston. Phage (~100 000 p.f.u.) were grown on *Escherichia coli* strain Q358 and duplicate sets of plaque filter lifts were prepared (Sambrook *et al.*, 1989). Each filter set was hybridized at reduced stringency with heterologous DNA probes comprising either coil 1a or tail domain sequences from *A.lumbricoides* IF cDNAs as described below. Positive plaques were purified by one or two additional rounds of screening and phage DNA was isolated from plate lysates. Phage were harvested in 50 mM Tris−HCl, pH 8.5, at 4°C, 10 mM $MgSO_4$ by rinsing overnight at 4°C using 7 ml buffer per 132 mm Petri dish. Bacterial debris was removed by low speed centrifugation and the supernatant then prewarmed to 37°C. DNase I and RNase A (Sigma) were each added to a final concentration of 1 μg/ml and incubation was continued at 37°C for 1 h. Nuclease treatment was terminated by the sequential addition of CDTA, proteinase K (Boehringer) and Na-sarcosyl to 20 mM, 100 μg/ml and 0.5% respectively, and subsequent incubation at 37°C for 1 h. The resulting crude phage DNA was precipitated with ethanol, resuspended in 500 μl of 50 mM Tris−HCl, pH 8.0, at 22°C, 10 mM CDTA, 0.5% Na-sarcosyl and finally purified by phenol extraction. Phage DNA digests generated with either *Eco*RI, *Hin*dIII, *Sac*I, *Sal*I, *Xba*I or *Xho*I (Gibco-BRL) were tested by Southern blot hybridization for subcloning of relevant genomic fragments. Subsequent to extensive restriction enzyme mapping suitably sized subfragments were ligated to M13mp18/19 vectors (Yanisch-Perron *et al.*, 1985) for preparation of single-strand sequencing templates. Sequencing was by the dideoxynucleotide chain termination method (Sanger *et al.*, 1977) using [α-$^{35}$S]dATPαS (Amersham) and Sequenase Version 2.0 (USB). Four *C.elegans* IF genes, CeIIF $a_1$, $a_2$, $b_1$ and $c_2$, were completely sequenced on both strands.

Three novel genes, CeIIF $a_3$, $a_4$ and $c_1$, were isolated from subgenomic plasmid libraries established with gel-purified genomic DNA fragments. The latter were detected by Southern blot hybridization at reduced stringency with coil 1a probes (see below) from the previously cloned four genes. Library constructions were as described (Dodemont *et al.*, 1990) with appropriately cut and dephosphorylated pUC18 (Yanisch-Perron *et al.*, 1985) as vector DNA and *E.coli* strain XL-1 Blue (Stratagene) as bacterial host. The eighth gene, CeIIF $b_2$, described in this study, was obtained by PCR-based cloning (see below). Characterization and sequence analysis of the new genes were as described above.

For each of the eight cloned IF genes at least one corresponding cDNA was isolated directly from a plasmid cDNA library. Additional cDNAs were obtained by RT-PCR on total mRNA (see below). The plasmid cDNA library representing total poly(A)$^+$ RNA from mixed-stage nematode populations was established using the XL-1 Blue/pUC18 host/vector system and comprised ~400 000 primary transformants. All methods dealing with cDNA synthesis, library construction and replication were as described (Dodemont *et al.*, 1990). The IF cDNAs were isolated by hybridization at high stringency to gene-specific probes specified below. All cDNAs were completely sequenced on both strands.

### *Hybridization of filter-bound nucleic acids*

Conditions for hybridization of Southern and Northern blots, 'polytene' YAC grid filters (a gift from Dr John Sulston), and replica filters containing lysed phage plaques or bacterial colonies were as follows. Hybridization at high stringency was preceded by prehybridization at 42°C for 4 h in 50% (v/v) formamide, 50 mM sodium phosphate, pH 6.8, 1 mM sodium pyrophosphate, 5 × SSC, 5 mM CDTA, 5 × Denhardt's solution, 0.5% SDS 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 sodium phosphate, 1 × Denhardt's solution and 5 ng/ml of total DNA probe (see below). After 15−20 h incubation at 42°C, filters were washed twice for 1 h at 42°C with hybridization solution without probe, followed by washes in 0.5 × SSC/0.1% SDS and 0.1 × SSC/0.1% SDS at 50−60°C for 15 min each. Hybridization at low stringency was as above except that at all stages the formamide concentration and incubation temperature were decreased to 25% (v/v) and 37°C. Filters were usually washed twice in 5 × SSC/0.1% SDS and if required once more in 0.5 × SSC/0.1% SDS at 55°C for 15 min each. Filters were autoradiographed to Kodak X-Omat AR film using intensifying screens.

### *Hybridization probes*

The first probes used in this study were derived from four full-length *A.lumbricoides* IF cDNAs. These were isolated from a plasmid library by hybridization at reduced stringency to a 174 bp *Pvu*II−*Acc*I fragment that comprised the coil 1a region from the cloned non-neuronal IF gene of the mollusc *H.aspersa* (Dodemont *et al.*, 1990). Details of the isolation and full characterization of the *Ascaris* IF cDNAs will be presented elsewhere. Two cDNAs, AscIF $B_1$ and AscIF $A_1$, encode the major muscle IF proteins B and A described previously (Weber *et al.*, 1989). The third cDNA, AscIF $A_2$, is highly related but not identical to AscIF $A_1$ whereas the fourth cDNA AscIF $C_1$ represents a separate class of IF-specific sequences. AscIF $B_1$ coil 1a probe: 142 bp *Alu*I−*Hae*III fragment covering amino acids 71−118 of the published B ($B_1$) protein sequence; amino acids (aa) indicated for the other *Ascaris* probes are relative to the $B_1$ sequence. AscIF $A_1$ coil 1a probe: 179 bp *Sau*3A−*Sau*3A fragment (aa 55−113); AscIF $A_2$ coil 1a probe: 219 bp *Taq*I−*Rsa*I fragment (aa 58−130); AscIF C1 coil 1a probe: 181 bp *Alu*I−*Rsa*I fragment (aa 63−126); AscIF $B_1$ tail domain probes comprising two adjacent fragments: 258 bp *Pst*I−*Bgl*II and 174 bp *Bgl*II−*Fok*I fragments (aa 450−589 plus 11 bp of the 3′-untranslated sequence); AscIF $A_2$ tail domain probes consisting of two partially overlapping fragments: 339 bp *Alu*I−*Alu*I and 178 bp *Bgl*II−*Pst*I fragments (aa 438−589 plus 15 bp of the 3′-untranslated sequence). The fragments described above were mixed at equimolar amounts to provide two separate hybridization cocktails of coil 1a and tail domain probes which enabled the isolation of the first four *C.elegans* IF genes $a_1$, $a_2$, $b_1$ and $c_2$ (see Results).

The following probes were derived from cloned *C.elegans* IF genes and comprised predominantly coil 1a sequences. Nucleotide (nt) designations listed below refer to the gene sequences deposited in the database, except for those of the $b_2$ probe which are given with respect to the corresponding cDNA sequence. CeIIF $b_1$ probe: 160 bp *Pst*I−*Hin*dIII fragment (nt 304−463); CeIIF $b_2$ probe: 186 bp PCR fragment (nt 159−344, see below); CeIIF $a_1$ probe: 168 bp *Dra*I−*Alu*I fragment (nt 2255−2422); CeIIF $a_2$ probe: 135 bp *Fok*I−*Fok*I fragment (nt 1524−1658); CeIIF $a_3$ probe: 192 bp *Sac*I−*Dde*I fragment (nt 359−550); CeIIF $a_4$ probe: 330 bp *Dra*I−*Sac*I fragment (nt 868−1197); CeIIF $c_1$ probe: 113 bp *Rsa*I−*Hin*dIII fragment (nt 1160−1272); CeIIF $c_2$ probe: 122 bp *Alu*I−*Rsa*I fragment (nt 1160−1281). These probes were used for isolation of novel IF genes as indicated above, for cDNA library screening and for hybridization of Northern blots and YAC grid filters.

Hybridization probes hp1 to hp7 used for Northern analysis (see Results) were derived from the CeIIF $c_2$ gene, except for hp6, which was isolated from the $c_2$-H cDNA. Hp1: ~2100 bp *Eco*RI−*Hin*dIII fragment (located

510 bp upstream to the ATG codon: nt 1076–1078); hp2: 122 bp *Alu*I–*Rsa*I fragment (nt 1160–1281); hp3: 219 bp *Rsa*I–*Rsa*I fragment (nt 1282–1500); hp4: 188 bp *Xho*I–*Sal*I fragment (nt 3173–3360); hp5: 261 bp *Sph*I–*Sac*I fragment (nt 3490–3750); hp6: 312 bp *Pvu*II–*Acc*I fragment (nt 3926–4521); hp7: ~5600 bp fragment (located 1100 bp downstream from two overlapping AATAAA polyadenylation signals of $c_2$-H RNA transcripts at nt 4566–4575). Nucleotide designations refer to the $c_2$ gene sequence; locations of the DNA fragments are given in Figure 9B.

All DNA fragments listed above were labeled by nick-translation (Rigby *et al.*, 1977) with [$\alpha$-$^{32}$P]dCTP (Amersham) to high specific activity (in excess of $10^9$ c.p.m./$\mu$g).

### PCR analysis and primers

Reaction mixtures (100 $\mu$l) for PCR (Saiki *et al.*, 1988) consisted of 50 mM Tris–HCl, pH 8.3, at 22°C, 50 mM KCl, 1.6 mM $MgCl_2$, 300 $\mu$M of each dNTP, 500 ng genomic DNA or 40 ng mRNA: cDNA hybrid, 400 ng (~0.5 $\mu$M) of each primer and 2.5 units of AmpliTaq DNA polymerase (Perkin-Elmer Cetus). Mixtures overlaid with mineral oil were heated at 95°C for 10 min and then carried through 40 cycles of denaturation (95°C, 1.5 min), annealing (55°C, 2 min) and polymerization (72°C, 2 min) which was followed by a final extension step of 10 min. Amplified DNAs were purified by preparative gel electrophoresis, treated with Klenow polymerase, kinased and ligated into the *Sma*I site of pUC/M13 vectors for subcloning or sequence analysis as described above.

Primers 1 and 2 were used for amplification of the 5′-region of the putative IF-specific EST, cm01g8 (Waterston *et al.*, 1992), from total mRNA by RT-PCR. Conditions for reverse transcription were as described (Dodemont *et al.*, 1990). Primers 1 (sense) and 2 (antisense) covered nt 10–32 and 195–173, respectively, of the EST sequence. The resulting single 186 bp PCR product was verified by sequence analysis and subsequently used for isolation of the corresponding full-length $b_2$ cDNA (see above). The latter served as a template for the design of additional primers 3–7, required for recovery of the $b_2$ gene. PCR on genomic DNA with primer pairs 3/4, 1/5 and 6/7 yielded three overlapping fragments comprising the 5′ (~2000 bp), middle (2272 bp) and 3′ portions (1591 bp) of the $b_2$ gene. Locations of the primers listed below are given with respect to the $b_2$-H cDNA sequence. Primer 1 (sense): 5′-GCAGCAGTTGAACTCTCGCCTTG-3′, nt 159–181; primer 2 (antisense): 5′-GAATGAGCTTCTGAGATCTCACG-3′, nt 344–322; primer 3 (sense): 5′-CGACTTCATCATGTCGGCGGTTAG-3′, nt 21–44; primer 4 (antisense): 5′-CTCCAATAGTTCCAGAATGAGC-3′, nt 358–337; primer 5 (antisense): 5′-CTTTCCTTGTGGATCGCACTCAACG-3′, nt 1356–1332; primer 6 (sense): 5′-GAGAGAGAAGCCAACACACTCCAG-3′, nt 982–1005; primer 7 (antisense): 5′-GGGAATGTAGCAAGTATGTACGTCG-3′, nt 1788–1764.

Antisense primers 8–10 were used in conjunction with sense primers that were identical to SL1 or SL2 leader sequences (Krause and Hirsh, 1987; Huang and Hirsh, 1989) to test for *trans*-splicing of $a_1$, $a_2$ and $a_3$ RNA transcripts. RT-PCR on total mRNA with the primer pair SL1/8 resulted in three distinct products corresponding to the 5′-ends of the $a_1$-H, $a_1$-M and $a_1$-L mRNAs with sizes of 454, 403 and 317 bp, respectively. The primer pair SL2/8 yielded the $a_1$-L mRNA specific 317 bp PCR product. RT-PCR on total mRNA with the primer combinations SL1/9 and SL1/10 led to the recovery of the 5′-ends of the $a_2$ and $a_3$ mRNAs, as 213 and 219 bp PCR products, respectively. Nucleotides for the antisense primers given below refer to their locations in the genes $a_1$ (primer 8), $a_2$ (primer 9) and $a_3$ (primer 10). Primers were SL1 (sense): 5′-GGTTTAATTACCCAAGTTTGAG-3′; SL2 (sense): 5′-GGTTTTAACCCAGTTACTCAAG-3′; primer 8 (antisense): 5′-GAACGAAGTGCGTCAAGATCGGCTG-3′, nt 2541–2517; primer 9 (antisense): 5′-GGACGCAGCATTCTGTCCGTATG-3′, nt 1534–1512; primer 10 (antisense): 5′-GGAAGCAGCGCCTTGGCCATATG-3′, nt 311–289.

## Acknowledgements

We thank Tomma Eisbein for skilful technical assistance and Dr James Adjaye for his contributions to PCR analysis. We are grateful to Dr John Sulston (MRC Laboratory of Molecular Biology, Cambridge, UK) for providing the genomic library and the YAC grid filter. We are indebted to Dr Michael Granato (Max Planck Institute for Biochemistry, Martinsried, Germany) for the ACEDB analysis. We thank Matthias Burba for his computer designed artwork, Irene Gajewski for photography and Gillian Paterson for typing the manuscript. N.L. and D.R. were supported by postdoctoral fellowships from the Max Planck Society. All gene and cDNA sequences reported above are available from the EMBL/GenBank under accession numbers X70830–X70836, X78553 and X78554.

## References

Agabian,N. (1990) *Cell*, **61**, 1157–1160.

Albertson,D.G. and Thomson,J.N. (1976) *Phil. Trans. R. Soc. Lond. Ser. B. Biol.*, **275**, 299–325.

Bader,B.L., Magin,T.M., Hatzfeld,M. and Franke,W.W. (1986) *EMBO J.*, **5**, 1865–1875.

Barstead,R.J. and Waterston,R.H. (1991) *J. Cell Biol.*, **114**, 715–724.

Bartnik,E. and Weber,K. (1987) *Eur. J. Cell Biol.*, **45**, 291–301.

Bartnik,E. and Weber,K. (1989) *Eur. J. Cell Biol.*, **50**, 17–33.

Bartnik,E., Osborn,M. and Weber,K. (1985) *J. Cell Biol.*, **101**, 427–440.

Bartnik,E., Osborn,M. and Weber,K. (1986) *J. Cell Biol.*, **102**, 2033–2041.

Bartnik,E., Kossmagk-Stephan,K., Osborn,M. and Weber,K. (1987) *Eur. J. Cell Biol.*, **43**, 329–338.

Bektesh,S., Van Doren,K. and Hirsh,D. (1988) *Genes Dev.*, **2**, 1277–1283.

Birnstiel,M.L., Busslinger,M. and Strub,K. (1985) *Cell*, **41**, 349–359.

Blumenthal,T. and Thomas,J. (1988) *Trends Genet.*, **4**, 305–308.

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

Coulombe,P.A., Hutton,M.E., Letai,A., Hebert,A., Paller,A.S. and Fuchs,E. (1991) *Cell*, **66**, 1301–1311.

Coulson,A., Sulston,J., Brenner,S. and Karn,J. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 7821–7825.

Coulson,A., Kozono,Y., Lutterbach,B., Shownkeen,R., Sulston,J. and Waterston,R. (1991) *BioEssays*, **13**, 413–417.

Dahlstrand,J., Zimmerman,L.B., McKay,R.D.G. and Lendahl,U. (1992) *J. Cell Sci.*, **103**, 589–597.

Dibb,N.J. and Newman,A.J. (1989) *EMBO J.*, **8**, 2015–2021.

Dodemont,H., Riemer,D. and Weber,K. (1990) *EMBO J.*, **9**, 4083–4094.

Döring,V. and Stick,R. (1990) *EMBO J.*, **9**, 4073–4081.

Fields,C. (1990) *Nucleic Acids Res.*, **18**, 1509–1512.

Fisher,D.Z., Chaudhary,N. and Blobel,G. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 6450–6454.

Francis,R. and Waterston,R.H. (1991) *J. Cell Biol.*, **114**, 465–479.

Fuchs,E. and Weber,K. (1994) *Annu. Rev. Biochem.*, **63**, 345–382.

Gounari,F., Merdes,A., Quinlan,R., Hess,J., FitzGerald,P.G., Ouzounis,C.A. and Georgatos,S.D. (1993) *J. Cell Biol.*, **121**, 847–853.

Huang,X.-Y. and Hirsh,D. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 8640–8644.

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.

Karn,J., Matthes,H.W.D., Gait,M.J. and Brenner,S. (1984) *Gene*, **32**, 217–224.

Kozak,M. (1991) *J. Cell Biol.*, **115**, 887–903.

Krause,M. and Hirsh,D. (1987) *Cell*, **49**, 753–761.

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.

Lane,E.B., Rugg,E.L., Navsaria,H., Leigh,I.M., Heagerty,A.H.M., Ishida-Yamamoto,A. and Eady,R.A.J. (1992) *Nature*, **356**, 244–246.

Lees,J.F., Shneidman,P.S., Skuntz,S.F., Carden,M.J. and Lazzarini,R.A. (1988) *EMBO J.*, **7**, 1947–1955.

Leung-Hagesteijn,C., Spence,A.M., Stern,B.D., Zhou,Y., Su,M.-W., Hedgecock,E.M. and Culotti,J.G. (1992) *Cell*, **71**, 289–299.

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.

Osman,M., Paz,M., Landesman,Y., Fainsod,A. and Gruenbaum,Y. (1990) *Genomics*, **8**, 217–224.

Palmiter,R.D. (1974) *Biochemistry*, **13**, 3606–3615.

Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature*, **263**, 211–214.

Pruss,R.M., Mirsky,R., Raff,M.C., Thorpe,R., Dowding,A.J. and Anderton,B.H. (1981) *Cell*, **27**, 419–428.

Quax,W., Vree Egberts,W., Hendriks,W., Quax-Jeuken,Y. and Bloemendal,H. (1983) *Cell*, **35**, 215–223.

Remington,S.G. (1993) *J. Cell Sci.*, **105**, 1057–1068.

Riemer,D., Dodemont,H. and Weber,K. (1991) *Eur. J. Cell Biol.*, **56**, 351–357.

Riemer,D., Dodemont,H. and Weber,K. (1992) *Eur. J. Cell Biol.*, **58**, 128–135.

Riemer,D., Dodemont,H. and Weber,K. (1993) *Eur. J. Cell Biol.*, **62**, 214–223.

Rigby,P.W.J., Dieckmann,M., Rhodes,C. and Berg,P. (1977) *J. Mol. Biol.*, **113**, 237–251.

Rosenbluth,J. (1967) *J. Cell Biol.*, **34**, 15–33.

Rothnagel,J.A., Dominey,A.M., Dempsey,L.D., Longley,M.A., Greenhalgh,D.A., Gagne,T.A., Huber,M., Frenk,E., Hohl,D. and Roop,D.R. (1992) *Science*, **257**, 1128–1130.

Saiki,R.K., Gelfand,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988) *Science*, **239**, 487–491.

Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.

Spieth,J., Brooke,G., Kuersten,S., Lea,K. and Blumenthal,T. (1993) *Cell*, **73**, 521–532.

Steinert,P.M. and Roop,D.R. (1988) *Annu. Rev. Biochem.*, **57**, 593–625.

Sulston,J. and Hodgkin,J. (1988) In Wood,W.B. (ed.), *The Nematode Caenorhabditis elegans*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 587–606.

Sulston,J. *et al.* (1992) *Nature*, **356**, 37–41.

Szaro,B.G., Pant,H.C., Way,J. and Battey,J. (1991) *J. Biol. Chem.*, **266**, 15035–15041.

Tomarev,S.I., Zinovieva,R.D. and Piatigorsky,J. (1993) *Biochim. Biophys. Acta*, **1216**, 245–254.

Tyner,A.L., Eichman,M.J. and Fuchs,E. (1985) *Proc. Natl Acad. Sci. USA*, **82**, 4683–4687.

Ward,S., Burke,D.J., Sulston,J.E., Coulson,A.R., Albertson,D.G., Ammons,D., Klass,M. and Hogan,E. (1988) *J. Mol. Biol.*, **199**, 1–13.

Waterston,R. *et al.* (1992) *Nature Genet.*, **1**, 114–123.

Way,J., Hellmich,M.R., Jaffe,H., Szaro,B., Pant,H.C., Gainer,H. and Battey,J. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 6963–6967.

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.

Wood,W.B. (ed.) (1988) *The Nematode Caenorhabditis elegans*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) *Gene*, **33**, 103–119.

Zewe,M., Höger,T.H., Fink,T., Lichter,P., Krohne,G. and Franke,W.W. (1991) *Eur. J. Cell Biol.*, **56**, 342–350.

*Received on February 17, 1994; revised on March 17, 1994*