The sequence of a cytoplasmic intermediate filament (IF) protein from the annelid *Lumbricus terrestris* emphasizes a distinctive feature of protostomic IF proteins

Marc Bovenschulte, Dieter Riemer, Klaus Weber*

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

Received 11 January 1995

Abstract The complete cDNA clone for a cytoplasmic intermediate filament (IF) protein from the annelid *Lumbricus terrestris* reported here, shows an extra 42 residues in the coil 1b subdomain of the central rod, as do the IF proteins from nematodes and molluscs. These extra six heptads are also present in all nuclear lamins but not in any known vertebrate cytoplasmic IF protein. Thus, it seems that protostomic metazoa conserve a lamin-like structural element in their cytoplasmic IF proteins, which was lost in the deuterostomic metazoan branch leading to the vertebrates.

Key words: Evolution; Intermediate filament; Lamin; Protostomia; Vertebrate

1. Introduction

The multigene family encoding the structural proteins of the cytoplasmic intermediate filaments (IF) comprises close to 50 different members in a mammal. All IF proteins share a central α-helical rod domain flanked by highly variable terminal domains. Exonic sequences, intron patterns and expression rules define 4 types of cytoplasmic IF proteins (types I to IV) in addition to the nuclear lamins. Even the structurally conserved rod domains of the various IF-types can sometimes share only 20 to 30% sequence identity [1,2]. The evolution of IF proteins is still poorly understood, since outside the vertebrates molecular knowledge of IF proteins is restricted essentially to only two invertebrate phyla, the molluscs and the nematodes [3–9]. The cytoplasmic IF proteins of these invertebrates are more closely related to nuclear lamins than are the IF proteins of vertebrates [3,4]. All have the characteristic lamin length of the coil 1b domain and nearly all contain a lamin-like segment in their C-terminal tail domains [3–9]. The latter is missing in two of the three squid neurofilament proteins, which arise from the same gene [7] and in two of the eight *Caenorhabditis elegans* IF proteins so far characterized [9]. Curiously, even the rod domains from IF proteins of molluscs and nematodes can be remarkably different in sequence.

This strong drift in IF sequences raises the question whether the lamin-like length of the coil 1b domain found in all cytoplasmic IF proteins of nematodes and molluscs is just a peculiarity of these two invertebrate phyla or whether it could reflect a more general property of protostomic animals. Since IF are only poorly documented by electron microscopy in arthropods [10,11], while nearly all cell types of annelids show a wealth of IF [12], we have begun an analysis of cytoplasmic IF proteins of the annelid *Lumbricus terrestris*. Here, we describe a full length cDNA clone which characterizes a cytoplasmic IF protein with a lamin-like length of the coil 1b domain and an unexpectedly high sequence homology with a neurofilament protein from the squid *Loligo pealei.*

2. Materials and methods

The body musculature of the annelid together with the attached epidermis was extracted with buffer containing Triton X-100 and with buffers of high and low ionic strength as described [4]. The final residue was solubilized in sample buffer and subjected to preparative SDS-PAGE. A polypeptide band around 70,000, which reacted with the monoclonal antibody IFA [13] in immuno blots, was recovered by electroelution and lyophilization. The protein was freed of SDS by acetone extraction and subjected to CNBr cleavage. Fragments separated by HPLC were characterized by automated gas-phase sequencing.

Total cellular RNA from the entire worm was isolated essentialy as described [14] and poly(A)$^+$ RNA was purified on oligo (dT)-cellulose (Type 7; Pharmacia, Uppsala, Sweden) by affinity chromatography [5]. The cDNA synthesis and transformations were performed exactly as described. About 60,000 bacterial colonies were grown on nitrocellulose filters, replicated and lysed [14]. Filters were screened at reduced stringency (25% formamide, 5 x SSC, 37°C) [15] with a nick-translated [16] cDNA fragment covering amino acid residues glycine-65 to lysine-371 of the *C. elegans* IF protein A-1 [5]. After two rounds of screening positive clones were isolated and mapped following standard techniques [17]. Both DNA strands of all subfragments were fully sequenced using Sequenase 2.0 (USB, Cleveland, OH, USA).

3. Results

A cytoskeletal residue from epidermal-muscular tissue of *Lumbricus terrestris* obtained by extraction with Triton X-100, followed by buffers of high and low ionic strength was used in immuno blotting with the murine monoclonal antibody IFA, which reacts with many but not all IF proteins [13,15,18]. A moderately abundant polypeptide of apparent molecular weight 70,000 reacted positively with the antibody (data not shown). This polypeptide was isolated by preparative SDS-PAGE. Fragments obtained by CNBr cleavage were separated by HPLC and characterized by automated sequencing. Some of the sequences obtained could be aligned with known invertebrate IF sequences. They seemed to arise from the coil 1a region and the N- and C-terminal parts of the coil 2 region. These partial protein sequences established the presence of more than one IF protein in the preparation and indicated a good homology with IF proteins from the nematode *C. elegans*, particularly those of the A type [9]. Thus, we decided against cloning the *Lumbricus* IF protein via PCR and instead used a

*Corresponding author. Fax: (49) (551) 201 578.

SSDI 0014-5793(95)00108-5

```
   1 CTGAGGTCTCAAACACCTTTTGGGAAAGTCGGAAAGGAAGGAGCAGAAATGGCGAGCAAGACATCAACGGAAAAGACGATCACCGTGACGGAGGAGGTCTCGCGGTATCGGCCGACCATC  120
   1                                                 M  A  S  K  T  S  T  E  K  T  I  T  V  T  E  E  V  S  R  Y  R  P  T  I     24

 121 CAAGGTCGTAACATCATCATCCAGCGCCGATCGCAGGACTCTGGGCCCTCATCGTCGTCTTACAGCAAGAGGGAGTCGATATCCCGCTATGGAATGCACCCTGCGATGAGCCCGAACGCC  240
  25 Q  G  R  N  I  I  I  Q  R  R  S  Q  D  S  G  P  S  S  S  S  Y  S  K  R  E  S  I  S  R  Y  G  M  H  P  A  M  S  P  N  A     64

 241 TATGTCACCATGTCGAACACCGGCGTGACGGCCGTCAAGGAGTCGCGAGAGCAGGAGAAGAAGGACATGCAGGACCTCAACGAGCGACTGGCCAACTACATCGAGAAGGTTCGCTTCCTC  360
  65 Y  V  T  M  S  N  T  G  V  T  A  V  K  E  S  R  E  Q  E  K  K  D  M  Q  D  L  N  E  R  L  A  N  Y  I  E  K  V  R  F  L    104

 361 GAGGCCCAGAACAGGAAGCTGGCCGACGAGCTGCTCAAGTTGAAGGCCAAATGGGGAAAGGAGACATCCCAGATCAAAGCCATGTACCAGGCGGAGCTGGATGAGGCGAGGAGACTGCTG  480
 105 E  A  Q  N  R  K  L  A  D  E  L  L  K  L  K  A  K  W  G  K  E  T  S  Q  I  K  A  M  Y  Q  A  E  L  D  E  A  R  R  L  L    144

 481 GACGATGCTGTGAAGGAGAAGTCCCGCATGGAGATCAGATTGGCCTCGAACGAGGAGATGATGGACGAACTGAGACAGAGACTTGAGGATGCTCTGAAGGATGCGGCCGATGCCAAGGAG  600
 145 D  D  A  V  K  E  K  S  R  M  E  I  R  L  A  S  N  E  E  M  M  D  E  L  R  Q  R  L  E  D  A  L  K  D  A  A  D  A  K  E    184

 601 AAGTTCGGGCACCAGAACCAGCAGCTGTCAGACTTGGAGGGTGAGGTCGGTCTTCTTAGGAGACGTCTGGCCAGCCTAGAATCCGAGAGGGACAAAGAAAAGGGTCTGATCAAGAAACTA  720
 185 K  F  G  H  Q  N  Q  Q  L  S  D  L  E  G  E  V  G  L  L  R  R  R  L  A  S  L  E  S  E  R  D  K  E  K  G  L  I  K  K  L    224

 721 CAGGACGCTCTCAACGCTACCAGCATGGATCTGGACAATGAGACGTTGTTGCACATCGATGCCGAGAACAGAAGACAGACGTTGGAAGAGGAACTGGAATTCCTCAAGGCTCTTCATGAA  840
 225 Q  D  A  L  N  A  T  S  M  D  L  D  N  E  T  L  L  H  I  D  A  E  N  R  R  Q  T  L  E  E  E  L  E  F  L  K  A  L  H  E    264

 841 CAGGAACTGAAAGAACTGCAGGCGTTGGCCTACAGAGATTCAACAGCCGAGAACCGTGAATACTGGAAGACAGAGATGGGTCAGGCTCTTCGTGAGATCCAGGAAGCCTACGATGACAAG  960
 265 Q  E  L  K  E  L  Q  A  L  A  Y  R  D  S  T  A  E  N  R  E  Y  W  K  T  E  M  G  Q  A  L  R  E  I  Q  E  A  Y  D  D  K    304

 961 ATGGATGTCATGAGGGGAGAACTGGAGACGTATTACAACCTCAAGCTGCAAGAGTTCCGTACAGGGGCCACTCGTAACAACATGGAGTCGGTTCACTTCAAGGAGGAAAGCAAACGTCTT 1080
 305 M  D  V  M  R  G  E  L  E  T  Y  Y  N  L  K  L  Q  E  F  R  T  G  A  T  R  N  N  M  E  S  V  H  F  K  E  E  S  K  R  L    344

1081 CGTGACCAGATGCAAGCCATGCGTGACAAACTCAACGATGCTGAGAACAAGTTGTGCGCAGTATCTCGCGAGTTGGATCAGCTGAGGCGAGAGAAGGAGGAGCGCGAGCGCGAGTTGGAG 1200
 345 R  D  Q  M  Q  A  M  R  D  K  L  N  D  A  E  N  K  L  C  A  V  S  R  E  L  D  Q  L  R  R  E  K  E  E  R  E  R  E  L  E    384

1201 CACAGGAACGGCGAGTTGTCCGACCAAGTGATCAAGCTGCAGGCTGAGATGGAGGCAATGCTCAGGGAACTCCAGATGATCATCGACGCCAAACTCGGACTTGAACTCGAAATCATCACC 1320
 385 H  R  N  G  E  L  S  D  Q  V  I  K  L  Q  A  E  M  E  A  M  L  R  E  L  Q  M  I  I  D  A  K  L  G  L  E  L  E  I  I  T    424

1321 TACCGTCGTCTGCTCGAGGGAGAGGAAAGTCGCACTGGTCTTCGTCAAATCACCGACAATCTTCTGAACAGTGAATCGAACGAATACACGATCAGACAGACGGAGTCTACATCCGGAGAC 1440
 425 Y  R  R  L  L  E  G  E  E  S  R  T  G  L  R  Q  I  T  D  N  L  L  N  S  E  S  N  E  Y  T  I  R  Q  T  E  S  T  S  G  D    464

1441 AATTCGATGAGGGTGAGCCAGATAATGAAGGGTGAAATGTCAGTGAAGACAACCTATCAGAAAAGCGCAAAAGGTCCGGTTGCCATTTACGAGTGTTCTCAAGACGGCAAAACCGTTGCC 1560
 465 N  S  M  R  V  S  Q  I  M  K  G  E  M  S  V  K  T  T  Y  Q  K  S  A  K  G  P  V  A  I  Y  E  C  S  Q  D  G  K  T  V  A    504

1561 CTGGAAAACACCGGAAGAAAGGATGAATTGAAGGGCAACTGGAGTCTGACTCGCAACATTGATGGAAAGGATGTTGCCACTTACAAGTTCAGCGATTCCTTTGTTCTACGACCAGGACAG 1680
 505 L  E  N  T  G  R  K  D  E  L  K  G  N  W  S  L  T  R  N  I  D  G  K  D  V  A  T  Y  K  F  S  D  S  F  V  L  R  P  G  Q    544

1681 AAAATCAAGGGTTGGGCGAAGGGAACAAGACCTTTCACAGGGGCATCAGGTGACATCGAATCGGACCAGAACTGGGAGTCGTCGCATATCATCACCAAACTGGTCAACCCTCTCGGAGAG 1800
 545 K  I  K  G  W  A  K  G  T  R  P  F  T  G  A  S  G  D  I  E  S  D  Q  N  W  E  S  S  H  I  I  T  K  L  V  N  P  L  G  E    584

1801 GATCGTGCAACGCACATTCAACAGACGAAGTACGCATAGAGAGAGAATCGCATGTTGACTTCTGTACAATGCTTGCAGACAGAAAGAGAGAGAGAGAGAGACAACTAGCGAAAATCGTCT 1920
 585 D  R  A  T  H  I  Q  Q  T  K  Y  A  *                                                                                      596

1921 CGTCGACTGCTCGTCAATGTATAATCTTAAACGTCAGGAAATGGAATGCATATTGATTGAACATGATTGCTGTTGATTGAGTCTATTTTTATAGGCTGAAGTTTGATTGATTGATTGATC 2040
2041 TCTGAAGGGGTCCAGACGATTAAATTTTTTATAGCTTGACTAATGCTTCTTATTTAAAACGCTTGATTGCGATTTAAATTCCTTTATTATCACGATGGAACATTTATTTCAATCTCTCGT 2160
2161 CTTCCTTAGAGATGAACTTTAGTAGAAATGATCGTAGTTATCATTGTTTTGTCAAACAAGGCAGTAGGTGTAAAGAATTATTACGTACCAGTATAGCCTGCTATTAGCTTTGTCATTGCA 2280
2281 TAGCCATCTGTTTGGTGGTCGACAGAATTTTGAATGTAATAAAGCATGTTCTTTTCAACGATTAAAAAAAAAAAAAAAAA                                         2360
```

Fig. 1. Nucleotide and predicted amino acid sequence of the *Lumbricus terrestris* cDNA encoding IF1. The ATG start codon and the TAG stop codon are given in bold letters and are underlined. The canonical polyadenylation signal [27] AATAAAA 20 nucleotides upstream of the poly(A) tail is also marked. The sequence has been submitted to the EMBL/GenBank under Accession Number X83734.

cloning strategy based on hybridization at lower stringency with a cDNA probe from the *C. elegans* A-1 cDNA which encodes the N-terminal 80% of the rod domain of this protein [9] .

A cDNA library representing total *Lumbricus* mRNA was screened with the *C. elegans* IF probe (see section 2). Approximately 60,000 primary transformants were hybridized at reduced stringency to the cloned cDNA fragment. Hybridization resulted in three strong signals. Restriction and Southern analysis of the purified plasmid cDNAs yielded inserts of 1.8 to 2.4 kb. The largest insert was selected for sequence analysis. The cDNA clone IF-1 contains a single reading frame of 1788 bp starting with the initiator ATG at positions 49 to 51. The TGA stop codon corresponds to nucleotides 1837 to 1839 (Fig. 1). The ATG start codon is part of a sequence context considered favourable for translational initiation as defined originally for vertebrate mRNAs [19] and the predicted protein sequence fits the apparent molecular weight observed in SDS-PAGE (see above). The open reading frame is flanked by 5′- and 3′-untranslated sequences of 48 and 524 bp respectively. The latter sequence ends with a poly (A) tract of 17 residues, 20 bp downstream of the canonical polyadenylation signal AATAAAA (positions 2318 to 2323).

The open reading frame of the cloned annelid cDNA (Fig. 1) predicts a protein with calculated values for the molecular weight and pI of 68,800 and 5.5 respectively. It is referred to as IF-1 since it shows all the hallmarks of an invertebrate cytoplasmic IF protein. It has the tripartate sequence organization of all IF proteins, in which a central α-helical domain, able to form a coiled coil, is flanked by non-helical terminal domains [1,2] and, in addition, shows a convincing sequence homology with other invertebrate IF proteins (Fig. 2). The C-terminal sequence of the rod domain is compatible with the reactivity of IF-1 with the murine antibody IFA since it lacks any of the amino acid replacements known to destroy the linear epitope [15,18]. Fig. 2 also shows that the annelid IF-1 protein contains an extra 42 residues in the coil 1b subdomain of the rod as do all previously characterized cytoplasmic IF proteins from nematodes and molluscs [3–9]. These extra six heptads are also present in the nuclear lamins of both vertebrates and invertebrates [20–24] but are lacking in all vertebrate cytoplasmic IF proteins [1,2].

```
                 ▼                              ▲              ▼
Lu 76 VKESREQEKKDMQDLNERLANYIEKVRFLEAQNRKLADELLKLKAKWGKETSQIKAMYQAELDEARRLLDDAVKEKSRME
Lo 89 VKANREREKQDMRDLNERFANYIEKVRFLEAQNKKLAGELEELKSKWGKETSAIKEMYETELEEARKLIDATNKEKITLD
Ce 79 IRDSREREKKEMSDLNDRLASYIEKVRFLEAQNRKLAADLDALRSKWGKDTHNIRNMYEGELVDAQKLIDETNKQRKDME
Ha 65 FRGTREKEKREMQNLNERLASYIEKVHFLDAQVKKLEAENEALRNRKSESLQPIRDAYENELAQARKVIDELSSTKGVSE

Lu    IRLASNEEMMDELRQRLEDALKDAADAKEKFGHQNQQLSDLEGEVGLLRRRLASLESERDKEKGLIKKLQDALNATSMDL
Lo    VRVTELIDQLERQQKDLEESRTYHQIDQEQIARQNQQLADLEGEISMLRRSIESLEKEKMRQSNILAKMNDEMEKMRMDL
Ce    GQLKKMQDELAEMRRKLEDATKGREQDRAKIDALLVTLSNLEAEISLLKRRIAQLEDEVKRIKQENQRLLSELQRARTDL
Ha    AKVAGLQDEIASLRELIVTYENQSKDYRKKIESLGNQIGEYEGELHTLRIRCGSLEDENAKVRELLDKIQEQNRRLRADL
                      __________________________________________

                                        ▲                      ▼
Lu    DNETLLHIDAENRRQTLEEELEFLKALHEQELKELQALAYRDSTAENREYWKTEMGQALREIQEAYDDKMDVMRGELETY
Lo    NNETINHLDAENRRQTLEEELEFQKDVHAQELKELAALAYRDTTAENREFWRNELAQAIRDIQQEYDAKCDQMRGDIEAY
Ce    DQETLNRIDYQNQVQTLLEEIDFLRRVHDNEIKELQTLASRDTTPENREFFKNELSSAIRDIREEYDQVNNVHRNDMESW
Ha    DTETAAHIEADCLAQTKTEEAEFYKDLLDQ-LELLKPEPIQIKGMDYAEFWKSELSKCVREIQSAYDEKIDMIQQDTEAK

Lu    YNLKLQEFRTGATRNNMESVHFKEESKRLRDQMQAMRDKLNDAENKLCAVSRELDQLRREKEERERELEHRNGELSDQVI
Lo    YNLKVQEFRTGATKQNMEVTRNKEENTKLKSNMTEIRNRLADLEARNAQLERTNQDLLRDLEEKDRQNELESCQYKEEIT
Ce    YRLKVQEIQTQSARQNMEQGYAKEEVKRLRTQLSDLRGKLADLESRNSLLEKQIQELNYQLEDDQRSYEAALNDRDSQIR
Ha    YSAQLNSLRSGNVKDGMQLQHVQEEVKKLRTQAGEKNAMYAELAAKFASLQAERDSIGRQCSELERELEELRIKYNQDIG

                                              ▲
Lu    KLQAEMEAMLRELQMIIDAKLGLELEIITYRRLLEGEESRT  436   —
Lo    KLRGEMESILKELQDLMDIKLSLELEIAAYRKLLEGEESRV  449   54 %
Ce    KMREECQALMVELQMLLDTKQTLDAEIAIYRKMLEGEENRA  439   45 %
Ha    DLSNELSAVLAQLQILTDAKITMELEIACYRKLLEGEESRV  424   37 %
```

Fig. 2. Sequence alignment of the rod domains of IF proteins from annelids, nematodes and molluscs. The central α-helical rod domains of the *Lumbricus terrestris* IF-1 protein (Lu, top line; from Fig. 1), the neurofilament protein NF70 from the squid *Loligo pealei* (Lo, second line) [6], the *C. elegans* IF-A1 protein (Ce, third line) [9] and the IF-A protein of the *Helix aspersa* (Ha, fourth line) [5] are aligned. Residue numbers refer to the first and last amino acids shown. The start and end of the subdomains of the rod [1] (coils 1a, 1b and 2) are indicated by arrowheads pointing down and up respectively. The six extra heptads in the coil 1b domain of these invertebrate IF proteins, which are absent in all vertebrate IF proteins, are underlined. Identical residues in the annelid sequence and any of the three other invertebrate sequences are marked by bold letters. The sequence identity levels in % versus the annelid sequence are given at end of the sequences. Note the unusually high sequence identity (54%) between the rod sequences of *Lumbricus terrestris* IF-1 and the neurofilament protein NF70 of the squid *Loligo pealei.*

The annelid IF-1 protein is surprisingly closely related to the neurofilament NF 70 protein from the squid *Loligo pealei* [6]. Over the rod domain the two proteins, which come from different metazoan phyla, display 54% sequence identify (Fig. 2). This value drops to 45 and 37% respectively when comparison is made with either the *C. elegans* IF-A1 protein [9] or the IF-A protein of *Helix aspersa* [5]. The squid NF-70 protein and the annelid IF protein are particularly closely related in the coil 1a region, in the linker between coils 1a and 1b, in the carboxyterminal part of coil 1b, in the linker between coil 1b and coil 2, and in the aminoterminal and carboxyterminal parts of coil 2 (see Fig. 2). The sequence homology also extends into the carboxyterminal tail domain following the rod. Fig. 3 shows the alignment of the last 110 residues which form the lamin tail homology segments earlier recognized in some cytoplasmic proteins from nematodes and molluscs [3–9]. Over this region the annelid and the squid IF proteins show 33 to 36% sequence identity with murine lamin B1 [24]. The sequences of the N-terminal head domains of invertebrate IF proteins are usually very diverse and differ strikingly in length [3–9]. In the case of the annelid and squid IF proteins the length difference is only 13 residues (see Fig. 2) and the two domains show a low but significant sequence homology (data not shown).

4. Discussion

The globular actin and tubulin molecules are very well conserved in eukaryotic evolution. In contrast, IF proteins, which are based on the coiled coil forming rod domain and strikingly different terminal domains, show strong sequence drifts already in the metazoa, the only eukaryotic kingdom for which molecular data on IF proteins exist. Except for the consensus sequences, located primarily at the ends of the rod, even this domain is more conserved in sequence principles than in actual sequences [1]. One advantage of this surprising sequence polymorphism may be its use in considering evolutionary questions and problems of metazoan phylogeny. The finding that cytoplasmic IF proteins from nematodes and molluscs conserve in

```
Mu La B1  437  SATGNVCIEEIDVDGKFIRLKNTSEQDQPMGGWEMIRKIGDTSVS-YKYT-SRYVLKAG
Lu IF     486  SAKGPVAIYECSQDGKTVALENTGRKDELKGNWSLTRNIDGKDVATYKFS-DSFVLRPG
Lo IF     506  TSKGSVSIKEADSQGCFIALE-TKKEENLTG-WKIVRKVDDNKV--YTYEIPNLVLKTG
                  ● ● ● ●    ●    ●  ●       ● ●   ●      ●  ●       ●●  ●

Mu La B1       QTVTVWAANAGVTASPPTDLIWKNQNSWGTGEDVKVILKNSQGEEVAQ--RSTVFK  547
Lu IF          QKIKGWAKGTRPFTGASGDI-ESDQN-WES-SHIITKLVNPLGEDRATHIQQTKYA  596
Lo IF          TVVKIWSKNHQAQAR-GDDLVSRENDTWGTGSNVVTILQNEKGEDKANYTQNTVYQ  615
                    ●            ●        ●         ● ●  ●●  ●     ●
```

Fig. 3. Lamin-tail homology segment in the tail domain of the *Lumbricus* IF-1 protein. The lamin-tail homology segments of the *Lumbricus* IF-1 protein (Lu IF; middle line; from Fig. 1) and the neurofilament protein NF70 from the squid *Loligo pealei* [6] (Lo IF; third line) are aligned with the corresponding tail region of murine nuclear lamin B1 [13] (Mu La B1; top line). Dashes indicate deletions introduced to optimize the alignment. Similar small deletions are also found in alignments of several lamins [28]. Residue numbers refer to the first and last amino acids shown. Identical residues in the lamin sequence and one or both of the cytoplasmic IF sequences are marked in bold letters. Dots mark positions in which all three sequences share the same residue. The lamin-tail homology segment of the cytoplasmic IF proteins extends to the C-terminal end, while the sequence of murine lamin B1 has an additional 40 residues.

contrast to vertebrate IF proteins the lamin-like length of the coil 1b domain and often display a lamin-homology segment in their tail domains has led to the hypothesis, that the archetype cytoplasmic IF protein arose from a mutated nuclear lamin lacking the nuclear location signal and the CaaX box [3–5,23]. A direct assessment of this view will, however, require the molecular characterization of IF proteins from additional eukaryotic kingdoms such as the protists, plants and fungi.

The long coil 1b version seen in all IF proteins of nematodes and molluscs [3–9] is now also documented for a third invertebrate phylum. The annelid IF protein characterized here also displays homology with lamins in its tail domain. Thus, it seems that the lamin-like length of the coil 1b domain found in cytoplasmic IF proteins from nematodes, molluscs and annelids may reflect a property common to all phyla from the protostomic branch of metazoa.

This leaves the question of the evolutionary origin of the shortened coil 1b domain present in all vertebrate IF proteins [1]. Since it is also found in the cephalochordate *Branchiostoma lanceolatum* [25] and in urochordates (D.R. and K.W., unpublished results) it most likely precedes the origin of the phylum of the chordates, which covers vertebrates, cephalochordates and urochordates. Chordates, echinoderms and hemichordates are the major phyla of the deuterostomic branch of metazoa [26]. Although no molecular information on IF proteins from the latter two phyla is yet available the most straightforward hypothesis involves the speculation that the short coil 1b domain arose already early in metazoan evolution, most likely with the origin of the deuterostomia. Subsequent gene duplication events would then have led to the precursors of the type I to IV IF proteins and to their multiple genes. The hypothesis, that the coil 1b length of IF proteins allows a distinction between protostomia and deuterostomia, can be tested by future cDNA cloning of IF proteins from echinoderms, hemichordates and from additional protostomic phyla.

References

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

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

[3] Weber, K., Plessman, U., Dodemont, H. and Kossmagk-Stephan, K. (1988) EMBO J. 7, 2995–3001.

[4] Weber, K., Plessmann, U. and Ulrich, W. (1989) EMBO J. 8, 3221–3227.

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

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

[7] 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.

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

[9] Dodemont, H., Riemer, D., Ledger, N. and Weber, K. (1994) EMBO J. 13, 2625–2638.

[10] Hirokawa, N. (1986) J. Cell Biol. 103, 33–39.

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

[12] Bartnik, E., Kossmagk-Stephan, K. and Weber, K. (1987) Eur. J. Cell. Biol. 44, 219–228.

[13] Pruss, R.M., Mirsky, R., Raff, M.C. Thorpe, R., Dowding, A.J. and Anderton, B.H. (1981) Cell, 27, 418–428.

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

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

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

[17] Sambrook. J Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning. A laboratory manual, 2nd edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

[18] Hatzfeld, M. and Weber, K. (1991) J. Cell Sci. 99, 351–362.

[19] Kozak, M. (1991) J. Cell Biol. 115, 887–903.

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

[21] Mc.Keon, F.D., Kirschner, M.W. and Caput, D. (1986) Nature 319, 463–468.

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

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

[24] Höger, T.T., Krohne, G. and Franke, W.W. (1988) Eur. J. Cell. Biol. 47, 283–290.

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

[26] Ruppert, E.E. and Barnes, R.D. (1994) Invertebrate Zoology, 6th edition, Saunders College Publishing, Harcourt Brace College Publishers; Fort Worth, Philadelphia, San Diego, New York, Orlando, San Antonio, Toronto, Montreal, London, Sydney, Tokyo.

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

[28] Bossie, C.A. and Sanders, M.M. (1993) J. Cell Sci. 104, 1263–1272.