The EMBO Journal vol.7 no.10 pp.2995 – 3001, 1988

Amino acid sequences and homopolymer-forming ability of the intermediate filament proteins from an invertebrate epithelium

Klaus Weber, Uwe Plessmann, Hubert Dodemont and Klaus Kossmagk-Stephan

Max-Planck-Institute for Biophysical Chemistry, D-34 Göttingen, FRG

Communicated by K.Weber

Intermediate filaments (IF) isolated from the oesophagus epithelium of the snail *Helix pomatia* contain two polypeptides of mol. wt 66 000 (A) and 52 000 (B), which we have now characterized by *in vitro* self-assembly studies and by protein sequences. A and B can each form morphologically normal IF and share extended regions of sequence identity. All A-specific sequences seem to locate to an extension of the carboxyl-terminal domain. Although the *Helix* protein(s) reveal the IF-consensus sequences at the ends of the coiled-coil, the remainder of the rod domain shows conservation of sequence principles rather than extended homology, when compared with any subtype of vertebrate IF proteins. Interestingly, the *Helix* proteins have the longer coil 1b domain found in nuclear lamins and not in cytoplasmic IF proteins of vertebrates. They lack, however, the karyophilic signal sequence typical for lamins. Obvious implications for IF evolution and structure are discussed.

Key words: desmin/epithelia/intermediate filaments/invertebrates/keratins/neurofilaments

Introduction

In mammals the multigene family of intermediate filaments (IF) includes more than 30 different proteins, which display cell and tissue-specific expression patterns (for recent reviews see Steinert *et al.*, 1985; Osborn and Weber, 1986). The keratin-subgroup, which is restricted to epithelia, is particularly complex. Some 19 keratins are known for human epithelia (Moll *et al.*, 1982) and an additional eight are restricted to the α-keratin filaments (hard keratins) specific for epidermal appendages such as hair and nails (Heid *et al.*, 1986; Lynch *et al.*, 1986). Morphologically and functionally distinct epithelia usually display different, but overlapping, sets of keratins, while changes in keratin composition occur in multilayered epithelia depending on the cellular environment. Non-epithelial cells have a much lower complexity of cytoplasmic IF. Of the three neurofilament proteins, characteristic for neurones, only one is a developmental marker (Shaw and Weber, 1982; Pachter and Liem, 1984). In addition vertebrates display glial-specific GFAP, myogenic desmin, mesenchymal vimentin and the recently discovered peripherin, which is primarily found in the peripheral nervous system (Portier *et al.*, 1984; Leonard *et al.*, 1988). Interestingly, epithelial keratins differ from all non-epithelial IF in an important structural aspect. Keratin IF are obligatory heteropolymers while the other IF are usually homopolymers. Keratin IF arise from equal numbers of keratin I- and II-type molecules, which are strikingly different in sequence (Moll *et al.*, 1982; Hanugoklu and Fuchs, 1983; Eichner *et al.*, 1984, 1986; Hatzfeld and Franke, 1985; Parry *et al.*, 1985; Quinlan *et al.*, 1985; Steinert *et al.*, 1985). Finally the nuclear lamins, for which a precise number is not yet known, also belong to the IF family (Aebi *et al.*, 1986; Fisher *et al.*, 1986; McKeon *et al.*, 1986; Krohne *et al.*, 1987).

The origins of the complexity of vertebrate cytoplasmic IF proteins in general, and of epithelial keratins in particular, are unknown. Since various mammalian cDNA probes used in hybridization studies on a few invertebrate species have not led to the corresponding invertebrate IF genes (Fuchs and Marchuk, 1983; Quax *et al.*, 1984; Lewis and Cowan, 1985) we have approached the problem of invertebrate IF proteins by more traditional means. Our immunological and biochemical results on molluscs, annelids and nematodes point to a lower IF complexity for invertebrates (Bartnik *et al.*, 1985, 1986, 1987a,b). Two IF prototypes could be distinguished. These are a neuronal and a non-neuronal prototype (Bartnik *et al.*, 1987a,b). The latter is characteristically found in all non-neuronal cells known to express IF by electron microscopic criteria. In all three phyla, non-neuronal IF contain two distinct but immunologically related polypeptides, whose molecular weights are often species dependent. In *Helix pomatia* the oesophagus epithelium is a rich source of non-neuronal IF. Purified filaments show two structural polypeptides with apparent mol. wts of 66 000 and 52 000 (Bartnik *et al.*, 1985), which we now refer to as A and B. Immunologically related IF proteins are present in all epithelia, including the epidermis, as well as in glial cells and in fibroblasts (Bartnik *et al.*, 1987a).

The early evolutionary divergence of neuronal and non-neuronal IF proteins raises an important question. Does the non-neuronal IF prototype of invertebrates follow the keratin-pairing principle or does it rather display properties expected for an evolutionary precursor of vimentin, desmin and GFAP? The latter non-epithelial IF proteins of vertebrates are particularly closely related in sequence (Quax *et al.*, 1984; Balcarek and Cowan, 1985; Geisler *et al.*, 1985; Geisler and Weber, 1986) and therefore presumably share a relatively recent common ancestor. Here we provide the first sequence of an invertebrate cytoplasmic IF protein. We show that A and B of the *Helix* oesophagus epithelium share extended regions of sequence identity and can both form homopolymeric IF.

Results

Both A and B form homopolymeric IF in vitro

Filaments purified from *Helix* oesophagus epithelium contain two polypeptides of mol. wt 66 000 (A) and 52 000 (B), which can be solubilized by 8 M urea. After removal of the

urea by dialysis spontaneous reconstitution of 10 nm filaments occurs (Bartnik *et al.*, 1985). Figure 1a shows that A and B solubilized in 8 M urea can be separated by cation-exchange chromatography on Mono S. The more acidic B protein eluted earlier than the more basic A protein (Figure 1b) in agreement with previous isoelectric focusing results (Bartnik *et al.*, 1985). Individual polypeptides were assayed for filament-forming ability by the three-step dialysis procedure used earlier for the mixture of A and B. Electron microscopy, performed after negative staining, revealed for each component morphologically normal IF with a diameter ~10 nm (Figure 2). Thus homopolymeric IF can be formed *in vitro* by either protein alone.

IF proteins A and B share extended regions of sequence identity

To understand the molecular relation between A and B the separated proteins were fragmented with CNBr. Reverse-phase high-pressure liquid chromatography (HPLC) gave very similar elution profiles. Automated gas phase sequencing showed that most CNBr fragments of A had their exact counterpart in B (Figure 3). To obtain a full sequence of the smaller B protein, CNBr fragments were enzymatically cleaved. Peptides separated on C18 were subjected to gas phase sequencing. Tryptic and chymotryptic digests of the intact protein provided overlaps for the CNBr fragments. The sequence proposed for B is very close to complete (Figure 3). It lacks most likely only two residues, a threonine and a serine, between the N-terminal blocking group and the lysine marked as residue 1 in Figure 3. We also cannot exclude the presence of a few additional residues at the C terminus, since a detailed C-terminal analysis was not done. The calculated mol. wt of B is 51 000, which compares well with the value of 52 000 obtained by gel electrophoresis.

The partial sequence of A (Figure 3) shows that all N-terminal sequences of the large CNBr fragments arising from the rod domain have their exact counterparts in the B sequence. In addition A and B show identical sequences along the head domains. The extra mass of A clearly arises from an extension of the C-terminal domain, which was characterized in sequence using a fragment obtained by cleavage at cysteine residue 410. Figure 3 shows that all sequences specific for A seem to form a unique extension from a shorter tail domain, which is shared by A and B. Gel electrophoresis and protein blotting after tryptophan cleavage of the proteins confirm this interpretation (data not shown). Both A and B provide a large fragment of mol. wt 32 000, which due to its blocking group must cover the N-terminal side (residues 1–271 in Figure 3). The second fragment of B, carrying the N-terminal sequence KSELSK has a mol. wt of 20 000. The same sequence is found in the case of A for two longer fragments in line with a C-terminal extension and the distribution of the additional tryptophan residues specific for the longer A tail (Figure 3). The calculated mol. wt of A is 63 000 and thus in fair agreement with the value of 66 000 obtained by gel electrophoresis. However, given a low yield peptide, the tail domain could extend by some 15 additional residues which are currently under study.

Sequence comparison with vertebrate IF proteins

The sequence of B shows the typical arrangement of an IF protein (Figure 3). A central rod domain is flanked by non-α-helical terminal domains (Geisler and Weber, 1982). Independent proof for such a protease-resistant domain is seen in Figure 1b. Mild chymotryptic treatment converts the *Helix* proteins into a fragment of 44 000 mol. wt, which is relatively resistant to further degradation. Along the rod domain the continuous system of heptads with primarily hydrophobic residues in *a* and *d* positions is interrupted by short non-α-helical spacers, which separate the coiled-coil-forming segments, coils 1a, 1b and 2 [for nomenclature of IF domains see Geisler and Weber (1982)]. The spacers display one and two proline residues respectively. With 360 residues the length of the rod exceeds the corresponding

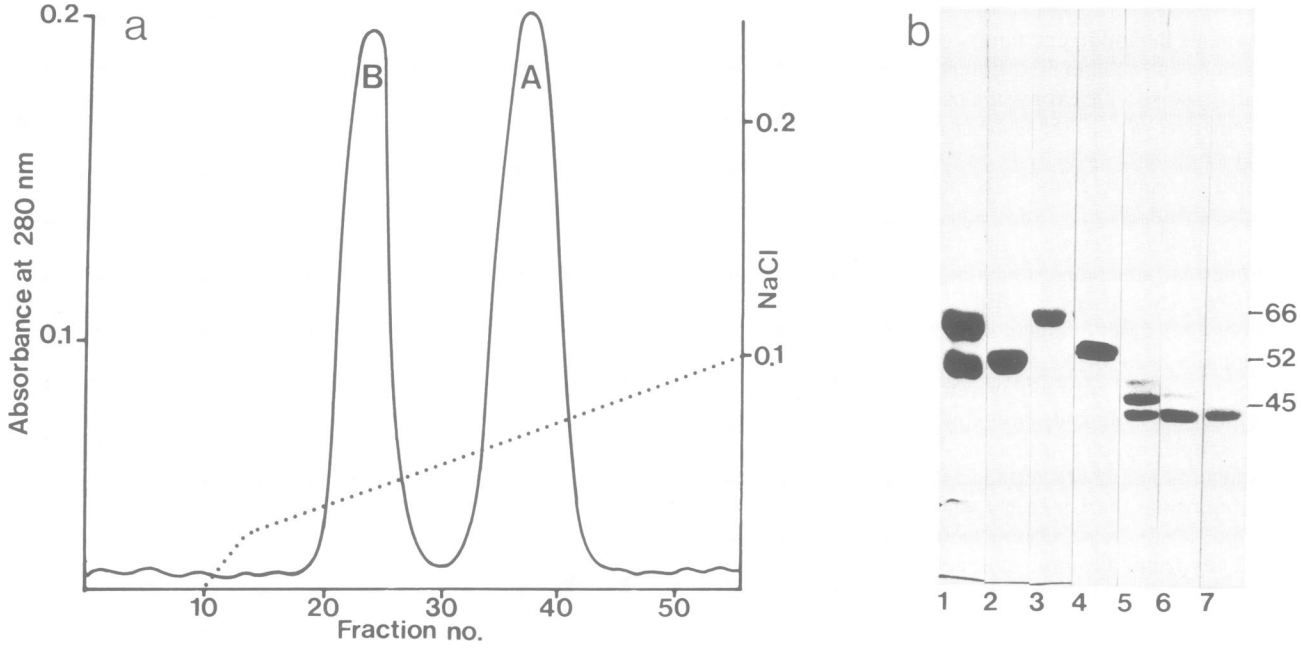

Fig. 1. Separation of IF proteins A and B (**a**) monitored by gel electrophoresis (**b**). IF purified from *Helix* oesophagus epithelium (**slot 1**) were dissolved in 8 M urea. Chromatography on Mono S (**a**) separated proteins B (**lane 2**) and A (**lane 3**). **Lanes 4–7** show the preparation of the protease-resistant rod domain of B obtained by mild chymotrypsin treatment for 0 (**lane 4**), 1 (**lane 5**), 3 (**lane 6**) and 5 (**lane 7**) min.

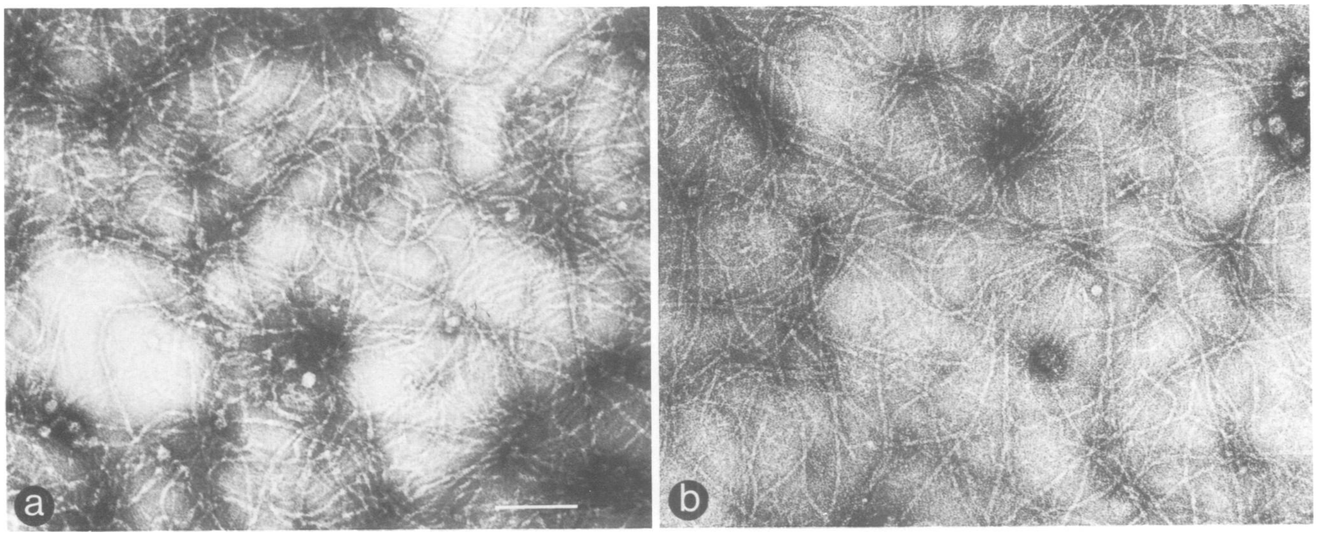

Fig. 2. *In vitro* filament formation by each of the two *Helix* proteins. After removal of urea by dialysis, protein A (**a**) and protein B (**b**) form morphologically normal IF. Micrographs are at the same final magnification. The bar in (a) indicates 0.1 μm.

```
    Head
    1
B   K I S T T Y E E E G R Q S K I Q P R A F V I T R S G P T S R S S S Y S A R Q S Y G S R S S I T P G V Y Q Q L S
A   K I S T T Y E E E G R Q S K I Q P R A F V I T R S G P T S R S S S Y S A R Q S Y G S R S S I T P G V Y Q Q L S

    56        Rod                      coil 1a
B   S S G I T D F R G T R E K E K R E M Q N L N E R L A G Y I E K V H F L D A Q V K K L E A E N E A L R N R K S E
A   S S G I T D F R G T R E K E K R E M Q N L N E R L A G Y I E K V H F L D A Q V K K L E A E N

    111             coil 1b
B   S L Q P I R D A Y E N E L A Q A R K V I D E L S S T K G V S E A K V A G L Q D E I A S L R E L I G T Y E N Q S

    166
B   K D Y R K K I E S L G N Q I G E Y E G E L H T L R I R C G S L E D E N A K V R E L L D K I Q D Q N R R L R A D

    221
B   L D T E T S A H I E A D C L A Q T K T E E C E F Y K D L L D Q L E L L K P E P M Q I K G M D Y A E F W K S
A                                                                                           M D Y A E F W K S

    coil 2
    274
B   E L S K C V R D I Q S A Y D E K I D I I Q Q D I E A K Y S S Q I N S L R S G N V K D G M Q L Q H V Q E E V K K
A   E L S K C V R D I Q S A Y D E K I D I I                                                 M Q L Q H V Q E E V K K

    329
B   L R T Q A G E K N A A Y A E L A A K F A S L Q A E R D A I G R Q C S E L E R E L E E L R I K Y N Q D I G D L S
A   L R T Q A G E K N A A Y A E

    384                                                                            Tail
B   N E L S A V L A Q L Q T L T D V K I T M E L E I A C Y R K L L E G E E S R V G L R S L V E Q A I G V Q G R G T
A                                         M E L E I A C Y R K L L E G E E S R V G L R S L V E Q A I G V Q G R G T

    439
B   A S L K D T I Q Q S
A   A S L K D T I Q Q S T A S G S M T V Q R S S K G P I A F N S V D Q S G S N I V I E N T T S G A R A K T Q S L K

    494
A   G W R V D K T V A G R V A A S I E L K N Y D L P P N T K Y T I W A K G A K D R A T A D N E Q I A D M F S L G V

    549
A   G
```

Fig. 3. Sequence of *Helix* IF protein B and the relation between proteins B and A. The B sequence shows the three domains of IF proteins (head, rod and tail). The coiled-coil segments along the rod (coils 1a, 1b and 2) are indicated by lines. Dots indicate the primarily hydrophobic residues present in the *a* and *d* positions of the consecutive heptads. For sequence comparison of the rod with a nuclear lamin see Figure 4. Current sequence data on A cover the head and tail domains and the N-terminal sequences of all large CNBr fragments from the rod. These are identical in sequence to the corresponding rod regions in B. The two proteins seem to differ only towards the C-terminal end where all A-specific sequences form an extension of the tail domain. Additional experiments on B, indicate that the N-terminal blocking group is separated by most likely only two residues from the lysine marked by 1. The precise order of the two residues is not yet known. For the possibility of a few additional residues at the C-terminal ends of both A and B, see text. For the blocking group see Note added in proof.

domain in various vertebrate cytoplasmic IF proteins. This is caused by a length increase of coil 1b by ~42 residues or six heptads as recently found for all nuclear lamins so far sequenced (Fisher *et al.*, 1986; McKeon *et al.*, 1986; Krohne *et al.*, 1987; Gruenbaum *et al.*, 1988). The two ends of the rod domain display the consensus-type sequences, which were already recognized when the first vertebrate IF proteins were compared (Geisler and Weber, 1982; Hanugoklu and Fuchs, 1983; Steinert *et al.*, 1983). They cover the N-terminal half of coil 1a and the 10 preceding residues (positions 63−91 in Figures 3 and 4) as well as the C-terminal region of coil 2 (positions 394−421 in Figures 3 and 4). Here the sequence YRKLLEGEXR is of particular interest. It is thought to contribute to the epitope of the monoclonal IFA antibody (Pruss *et al.*, 1981; Geisler *et al.*, 1983), which recognizes many but not all IF proteins (Bartnik *et al.*, 1987a; Magin *et al.*, 1987). Its reactions on *Helix* oesophagus led us previously to non-neuronal IF of invertebrates (Bartnik *et al.*, 1985).

```
63  RGTREKEKREMQNLNERLAGYIEKVHFLDAQVKKLEAENEALRNRKSESL
25  RITRLQEKEDLQELNDRLAVYIDRVRSLETENAGLRLRITESEEVVSREV

113 QPIRDAYENELAQARKVIDELSSTKGVSEAKVAGLQDEIASLRELIGTYE
75  SGIKAAYEAELGDARKTLDSVAKERARLQLELSKVREEFKELKARNTKKE

163 NQSKDYRKKIESLGNQIGEYEGELHTLRIRCGSLEDENAKVRELLDKIQD
125 GDLIAAQARLKDLEALLNSKEAALSTALSEKRTLEGELHDLRGQVAKLEA

213 QNRRLRADLDTETSAHIEADCLAQTKTEECEFYKDLLDQLELLKPEPMQI
175 ALGEAKKQLQDEMLRRVDAENRLQTMKEELDFQKNIYSEELRETKRRHET

263 KGMDYAEF....WKSELSKCVRDIQSAYDEKIDIIQQDIEAKYSSQINSL
225 RLVEIDNGKQREFESRLADALQELRAQHEDQVEQYKKELEKTYSAKLDNA

309 RSGNVKDGMQLQHVQEEVKKLRTQAGEKNAAYAELAAKFASLQAERDAIG
275 RQSAERNSNLVGAAHEELQQSRIRIDSLSAQLSQLQKQLAAKEAKLRDLE

359 RQCSELERELEELRIKYNQDIGDLSNELSAVLAQLQTLTDVKITMELEIA
325 DSLARERDTSRRLLAEKEREMAEMRARMQQQLDEYQELLDIKLALDMEIH

409 CYRKLLEGEESRV 421
375 AYRKLLEGEEERL 387
```

Fig. 4. Sequence alignment of the rod domain of *Helix* B protein and the corresponding region of human nuclear lamins A/C. The sequence of the *Helix* B protein (**upper line**) is from Figure 3 (residues 63−421). Human lamins A and C have identical rod domains. The lamin sequence (residues 25−387; **lower line**) is from Fisher *et al.* (1986) and McKeon *et al.* (1986). Note the single gap of four residues, indicated by points, in the B sequence. As seen in Figure 3 this gap locates to the region preceding coil 2. Arrowheads pointing down or up mark the starts and ends of coils 1a, 1b and 2. Identical or highly related residues (E = D, K = R = H, exchange of large hydrophobic residues and exchange of large aromatic residues) in the two sequences are marked by vertical lines. Both ends of the rods display the IF consensus sequences. Over the remainder of the domain sequence principles are conserved and homologous sequences are rare (see text).

A search for homologous proteins in the protein data bank detected all vertebrate IF proteins present in the bank. The preferred alignment used the human nuclear lamins A/C and is shown in Figure 4. It requires the introduction of only one gap, which corresponds to four additional residues of lamin A/C in the spacer between coils 1b and 2. This region is already known from earlier studies on other IF proteins to tolerate some changes in sequence length (see for instance Geisler *et al.*, 1984; Geisler and Weber, 1986; McKeon *et al.*, 1986). Figure 4 strongly emphasizes the consensus sequences at both ends of the rod. It also shows, however, that over the remainder of the domain sequence principles rather than actual sequences are conserved. The C-terminal tail domains of the *Helix* proteins show a moderate sequence homology with the larger tail domains of the nuclear lamins (Figure 5). This homology concerns the N-terminal part of the lamin tails and is particularly evident around the two tryptophans of the tail domain of the *Helix* A protein. Interestingly in this alignment the region surrounding the karyophilic motif of the nuclear lamins (Fisher *et al.*, 1986; McKeon *et al.*, 1986; Krohne *et al.*, 1987) is deleted in the cytoplasmic IF proteins and does not occur at any other place of the sequence.

The N-terminal head domain (Figure 3) is non-α-helical as in vertebrate IF proteins and shows the previously recognized sequence hypervariability (Geisler and Weber, 1982, 1986). Interestingly it lacks the ordered tracks of glycines and serines flanked by large hydrophobic residues, which are found in many keratins (see for instance Hanugoklu and Fuchs, 1983; Steinert *et al.*, 1983; Bader *et al.*, 1986). The presence of six arginines parallels the corresponding domain of many vertebrate non-epithelial IF proteins (Geisler and Weber, 1986), but the cluster of three glutamic acid residues close to the N terminus is unusual. The head domain contains the sequence RXS four times (positions 11−13, 30−32, 37−39 and 43−45; see Figure 3). In the same domain of several vertebrate IF proteins this

```
     390
XIL1 LSPSP---SRVTVSRASSSRAVR---TTKGKRKRIDVEESEASSSVSIDHSAAATGDVSIEEVDVDGK
XIA  LSPSPNTQKRSARTIASHSGAHI--SSSASKRRRL--EEGE-SRSSSFTQHARTTGKVSVEEVDPEGK
HsA  LSPSPTSQ-RS-RGRASSHSSQTQGGGSVTKKRKL--ESTE-SRS-SFSQHARTSGRVAVEEVDEEGK
A    LR-SLVEQAIGVQGRGTASLKDTIQQS---------------TASGSMTVQRSSKGPIAFNSVDQSGS
     423

                                                              505
XIL1 YIRLKNNSE----KDHPLGGWELTRTIG-EASVNFKFTSRYVLKAEQTVTIWAADAGVK
XIA  YVRLRNKSN----EDQSLGNWQIKRQIGDETPIVYKFPPRTLLKAGQTVTIWASGAGAT
HsA  FVRLRNKSN----EDQSMGNWQIKRQNGDDPLLTYRFPPKTFLKAGQVVTIWAAGAGAT
A    NIVIENTTSGARAKTQSLKGWRVDKTVAGRVAASIEL--NDYLPPNTKYTIWAKGAKDR
                                                              532
```

Fig. 5. Sequence alignment of the N-terminal part of the tail domain of the three nuclear lamins and the corresponding region of the larger *Helix* IF protein A. Alignment of two *Xenopus* lamins XIL1 and XIA and the human lamin A (HsA) follows essentially the presentation of Krohne *et al.* (1987). Bold letter mark identical residues in the *Helix* protein and in one or more lamin molecules. The karyophilic signal region of the nuclear lamins, marked by triangles above the first line, is deleted in the *Helix* cytoplasmic IF protein.

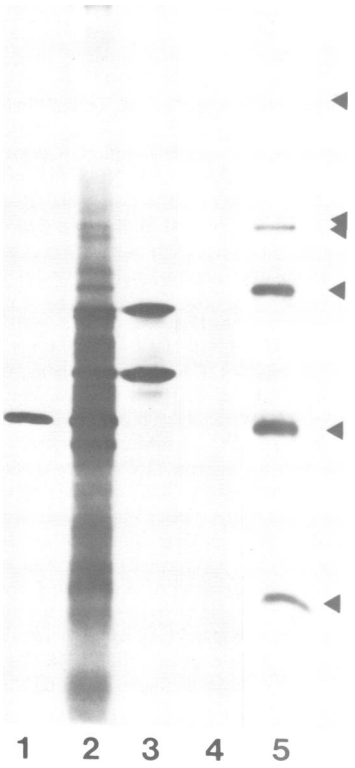

Fig. 6. Immunoprecipitation analysis of *Helix* IF polypeptides synthesized *in vitro*. Total mRNA from oesophagus epithelial cells was translated *in vitro* in a rabbit reticulocyte lysate, supplemented with L-[^{35}S]methionine. Newly synthesized polypeptides were subjected to SDS–gel electrophoresis either directly (**lane 2**) or after immunoprecipitation with rabbit antibodies to *Helix* IF proteins (**lane 3**) or the corresponding preimmune serum (**lane 4**). Blank incubation of the cell-free system and radiocatively labelled marker polypeptides (mol. wts 200, 100, 92.5, 69, 46 and 30 kd) are shown in **lanes 1** and **5** respectively. The fluorograph of a 4-day exposure to Kodak XAR-5 film is shown.

motif is recognized by protein-kinase A (Geisler and Weber, 1988). An additional motif occurs in the tail domain of the A protein (positions 458–460).

In vitro synthesis of Helix proteins

Although IF were always prepared in the presence of various protease inhibitors (Bartnik *et al.*, 1985; see Materials and methods) the sequence data in Figure 3 do not exclude a possible proteolytic derivation of B from A. Therefore total mRNA from *Helix* oesophagus epithelium was translated in a reticulocyte system. Immunoprecipitates obtained with a rabbit antibody known to react with A and B (Bartnik *et al.*, 1987a) reveal both components (Figure 6). Thus a simple proteolytic derivation of B from A seems unlikely.

Discussion

Vertebrate epithelial keratin filaments arise as obligatory heteropolymers from two distinct subtype molecules of quite distinct rod sequences. Keratins I are smaller and more acidic than keratins II (for references see Introduction). Since IF isolated from the *Helix* oesophagus epithelium contain components A (66 kd; pI 6.4) and B (52 kd; pI 6.0) in nearly equal amounts (Bartnik *et al.*, 1985) it was tempting to think of the epithelial *Helix* proteins as keratin-analogues. This is clearly not the case. We have shown that both A and B can form homopolymeric IF *in vitro*. The nearly complete sequence of B and the extended sequence data on A show that the two proteins probably differ only at the carboxyl end. Here all A specific sequences form an extension from a common tail domain. The basic character of this extension explains the separation of A and B in isoelectric focusing (Bartnik *et al.*, 1985) and ion exchange chromatography (Figure 1).

The large region of shared amino acid sequences also explains earlier immunological results on non-neuronal IF proteins in three invertebrate phyla. Regardless of the antigen used, polyclonal as well as most monoclonal antibodies reacted with both A and B. The few polypeptide-specific antibodies always involved the higher molecular weight species of the two proteins (Bartnik *et al.*, 1987a,b). Current protein sequence data (Figure 3) cannot exclude a proteolytic derivation of B from the longer A molecule. However, experiments using total mRNA of *Helix* oesophagus show that both A and B are synthesized *in vitro* by a reticulocyte system in a ratio which seems to preclude a precursor–product relationship (Figure 6). Additional experiments with cDNA clones are needed to decide whether A and B reflect the products of different genes or arise from a common transcript due to differential mRNA splicing.

Unexpectedly the *Helix* cytoplasmic IF proteins differ from all their vertebrate counterparts by a length increase of the coil 1b domain. This increase by ~42 residues or six heptads seems also a characteristic feature of the nuclear lamins (Fisher *et al.*, 1986; McKeon *et al.*, 1986; Krohne *et al.*, 1987; Gruenbaum *et al.*, 1988). Interestingly the long tail domain of the *Helix* A protein shows a moderate homology with the lamin tail domains. While this homology could be of evolutionary significance (Osborn and Weber, 1986; Myers *et al.*, 1987) it does not imply a nuclear lamin function. As shown in Figure 5 the region covering the karyophilic signal, thought to dictate lamin entry into the nuclear compartment (Fisher *et al.*, 1986; McKeon *et al.*, 1986; Krohne *et al.*, 1987; Gruenbaum *et al.*, 1988) is deleted in the tail domain of A and does not occur at any other place of the sequence.

Our sequence data on the rod domain of the oesophagus epithelium B protein have direct implications for future IF structure. They emphasize an amazing tolerance for residue changes over much of the presumptive coiled-coil as long as the sequence principles of the domain are not perturbed, i.e. the consecutive heptad patterns and the location of the non-α-helical spacers. Therefore the striking conservation of the consensus sequences, at both ends of the rod domain, indicate that the actual sequences present play an important, but not yet understood, part in the structure and/or function of IF. As shown earlier for vertebrate IF proteins these two consensus sequences cover the N-terminal part of the rod, extending about halfway through coil 1a, and the carboxyl end of coil 2 (Geisler and Weber, 1982, 1986; Hanugoklu and Fuchs, 1983; Steinert *et al.*, 1985).

It is not yet known which particular residues or sequence motifs along the rod provide a vertebrate IF protein with homopolymer-forming ability or dictate the need to find a complementary partner as in the keratins. This problem has also not been solved with the sequence of the *Helix* B protein. The optimal alignment given by the computer program, i.e. the relation of B with a nuclear lamin (Figure 4), reflects primarily the presence of a similarly large coil 1b domain

and not a particularly high sequence homology in this region. If only identical residues are counted the rod of B shows some 20–30% identity with the corresponding region of most currently known vertebrate IF proteins. Thus at least from linear sequence analysis the rod of B provides no convincing clue for the homopolymer-forming ability of the protein established by electron microscopy. We note, however, a peculiarity of residues 61–130, i.e. the region extending from the N-terminal end of the rod to a position some 10 residues past the start of coil 1b. Along this stretch the B protein seems more closely related to all vertebrate non-epithelial IF proteins than to the various keratins present in the data bank. The importance of this higher homology, if any, is unknown. It separates, however, the B protein not only from the keratins but also from the lamins, which are thought to form homopolymeric structures as far as the participating rod sequences are concerned (Aebi *et al.*, 1986). Maybe more vertebrate IF sequences will allow evolutionary drift to be distinguished from true and distinct structural requirements of the different branches of the IF protein family.

Although we cannot yet pinpoint the sequence features responsible for homo- and heteropolymer formation of IF proteins, the *Helix* proteins resemble by self-assembly criteria the vertebrate homopolymer-forming proteins and not the vertebrate epithelial keratins. This raises two questions. How general are our results on *Helix* for invertebrates and when did the keratin-pairing principle arise in metazoan evolution? Previous immunological results showed that A- and B-type polypeptides are found by blotting experiments in a variety of *Helix* epithelia including the epidermis and in glia cells. The same antibodies also reacted in immunofluorescence microscopy on fibroblasts (Bartnik *et al.*, 1985, 1987a). Thus at least in gastropods the immunologically defined non-neuronal IF type (Bartnik *et al.*, 1987a,b) involves either the same or several distinct but highly homologous molecules. Given our results on IF from the oesophagus epithelium, we would expect that such IF do not display the keratin-pairing principle. Similar immunological studies on annelids and the large nematode *Ascaris* emphasize again a non-neuronal IF prototype present in all epithelia as well as those non-neuronal cells, which by electron microscopical criteria express IF (Bartnik *et al.*, 1986, 1987b). Thus the rigorous separation of epithelial IF (keratins) from non-epithelial-non-neuronal IF documented for vertebrates has not yet emerged in those invertebrate phyla, which have been analysed at least by antibody reactivity patterns. It seems widely assumed (see for instance Guidice and Fuchs, 1987) that the high complexity of vertebrate keratins (Moll *et al.*, 1982) and their heteropolymer character offer advantages for gene-regulation mechanisms governing the expression of a structure. This situation may be exploited in transitional and stratified epithelia, where expression of different keratins can be influenced by cell position and differentiation. As invertebrates essentially lack such complex epithelia and are characterized by simple epithelia a provocative speculation arises. Did the keratin heteropolymer principle evolve essentially parallel to the acquisition of complex epithelia by the chordate/vertebrate lineage?

Materials and methods

Oesophagus IF from *H.pomatia* were purified as before (Bartnik *et al.*, 1985) Final pellets were suspended in 5 mM Tris–HCl, pH 7.4, 5 mM in ATP, incubated for 15 min at 4°C, harvested by centrifugation and stored in 50% glycerol at −20°C. All buffers contained the protease inhibitors previously given and in addition were made 5 μM in E64 (Peptide Institute, Osaka, Japan). Pellets were dissolved in freshly prepared urea–phosphate buffer (8 M urea, 5 mM 2-mercaptoethanol, 20 mM sodium phosphate, final pH 6.6) and dialysed for 2 h against the same buffer. Solutions clarified by centrifugation were applied to a Mono S column (Pharmacia) equilibrated with urea–phosphate buffer. Gradient elution used the same buffer and a urea–phosphate buffer containing 0.4 M NaCl (final pH 6.35) on a FPLC apparatus (Pharmacia). Pooled fractions of A and B were dialysed into urea–Tris buffer and then subjected to the three-step dialysis procedure to remove the urea (Bartnik *et al.*, 1985). Reconstitution of IF was monitored by electron microscopy after negative staining with 1% uranyl acetate.

For protein chemical studies A and B were reacted with vinyl-pyridine to modify the cysteine residues (Friedman *et al.*, 1970), dialysed against water and lyophilized. After CNBr cleavage, fragments were separated by HPLC on C4 or C18 columns. Gas phase sequencing was on an Applied Biosystems sequenator (model A470) with an on-line PTH-amino acid analyser. For mild chymotryptic digestion component B (0.15 mg/ml) was dialysed into 10 mM NH_4HCO_3, 1 mM 2-mercaptoethanol and treated at room temperature with the enzyme (1:250, w/w). Aliquots removed at different times were analysed by SDS–gel electrophoresis (Figure 1). Enzymatic digests obtained by standard procedures were separated on a C18 column. Purified peptides were sequenced as above. Methionine-containing tryptic and chymotryptic peptides were obtained from a derivative of B. After reaction of the cysteine residues (see above) the protein was treated with [^{14}C]iodoacetic acid to specifically label methionine residues (Platt *et al.*, 1973). Cleavage at cysteine with 2-nitro-5-thiocyanobenzoate (NTCB) was as before (Geisler *et al.*, 1982). Resulting fragments were subjected to ion exchange chromatography on Mono S as described above for the intact proteins.

Total poly(A)-containing polysomal RNA was isolated as described by Ramaekers *et al.* (1982). Translations *in vitro* were performed in a nuclease-treated rabbit reticulocyte lysate in the presence of L-[^{35}S]methionine (Amersham). For immunoprecipitation analysis aliquots were diluted 40-fold in a buffer containing 50 mM Tris–HCl, 150 mM NaCl, 5 mM EDTA, 5 mM EGTA, 10 mM methionine, 1% sodium deoxycholate, 1% Triton X-100, 0.2% sodium sarcosyl, final pH 8.5 at 4°C. After overnight incubation samples were treated with pre-immune rabbit IgGs followed by adsorption to *Staphylococcus aureus* cells (BioMakor). Pre-adsorbed samples were incubated at 4°C for 2 h with either affinity-purified rabbit antibodies to *H.pomatia* IF proteins (Bartnik *et al.*, 1987a) or the corresponding pre-immune serum. Immunoprecipitates were adsorbed to *S.aureus* cells, thoroughly washed in the incubation buffer supplemented with 0.025% 2-mercaptoethanol, boiled in sample buffer and subjected to electrophoresis in SDS–10% polyacrylamide gels, which were processed for fluorography.

Acknowledgements

We thank Dr M.Osborn for discussions, K.K.-S. was supported by a Max-Planck postdoctoral fellowship.

References

Aebi,U., Cohn,J., Buhle,L. and Gerace,L. (1986) *Nature*, **323**, 560–564.
Bader,B.L., Magin,T.M., Hatzfeld,M. and Franke,W.W. (1986) *EMBO J.*, **5**, 1865–1875.
Balcarek,J.M. and Cowan,N.J. (1985) *Nucleic Acids Res.*, **13**, 5527–5543.
Bartnik,E., Osborn,M. and Weber,K. (1985) *J. Cell Biol.*, **101**, 427–440.
Bartnik,E., Osborn,M. and Weber,K. (1986) *J. Cell Biol.*, **102**, 2033–2041.
Bartnik,E., Kossmagk-Stephan,K., Osborn,M. and Weber,K. (1987a) *Eur. J. Cell Biol.*, **43**, 329–338.
Bartnik,E., Kossmagk-Stephan,K. and Weber,K. (1987b) *Eur. J. Cell Biol.*, **44**, 219–228.
Eichner,R., Bonitz,P. and Sun,T.T. (1984) *J. Cell Biol.*, **98**, 1388–1396.
Eichner,R., Sun,T.T. and Aebi,U. (1986) *J. Cell Biol.*, **102**, 1767–1777.
Fisher,D., Chaudhary,N. and Blobel,G. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 6450–6456.
Friedman,M., Krull,L.H. and Cavins,J. (1970) *J. Biol. Chem.*, **245**, 3868–3871.
Fuchs,E. and Marchuk,D. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 5857–5861.
Geisler,N. and Weber,K. (1982) *EMBO J.*, **1**, 1649–1656.
Geisler,N. and Weber,K. (1986) In Shay,J.W. (ed.), *Cellular and Molecular Biology of the Cytoskeleton*. Plenum Press, New York, pp. 41–68.
Geisler,N. and Weber,K. (1988) *EMBO J.*, **7**, 15–20.
Geisler,N., Kaufmann,E. and Weber,K. (1982) *Cell*, **30**, 277–286.

Geisler,N., Kaufmann,E., Fischer,S., Plessmann,U. and Weber,K. (1983) *EMBO J.*, **2**, 1295–1302.
Geisler,N., Fischer,S., Vandekerckhove,J., Plessmann,U. and Weber,K. (1984) *EMBO J.*, **3**, 2701–2706.
Geisler,N., Plessmann,U. and Weber,K. (1985) *FEBS Lett.*, **182**, 475–478.
Giudice,G.J. and Fuchs,E. (1987) *Cell*, **48**, 453–463.
Gruenbaum,Y., Landesman,Y., Drees,B., Bare,J.W., Saumweber,H., Paddy,M.R., Sedat,J.W., Smith,D.E., Benton,B.M. and Fisher,P.A. (1988) *J. Cell Biol.*, **106**, 585–596.
Hanugoklu,I. and Fuchs,E. (1983) *Cell*, **33**, 915–924.
Hatzfeld,M. and Franke,W.W. (1985) *J. Cell Biol.*, **101**, 1826–1841.
Heid,H.W., Werner,E. and Franke,W.W. (1986) *Differentiation*, **32**, 101–119.
Krohne,G., Wolin,S.L., McKeon,F.D., Franke,W.W. and Kirschner,M.W. (1987) *EMBO J.*, **6**, 3801–3808.
Leonard,D.G.B., Gorham,J.D., Cole,P., Greene,L.A. and Ziff,E.B. (1988) *J. Cell Biol.*, **106**, 181–193.
Lewis,S.A. and Cowan,N.J. (1985) *J. Cell Biol.*, **100**, 843–850.
Lynch,M.H., O'Guin,W.M., Hardy,C., Mak,L. and Sun,T.. (1986) *J. Cell Biol.*, **103**, 2593–2606.
Magin,T.M., Hatzfeld,M. and Franke,W.W. (1987) *EMBO J.*, **6**, 2607–2615.
McKeon,F.D., Kirschner,M.W. and Caput,D. (1986) *Nature*, **319**, 463–468.
Moll,R., Franke,W.W., Schiller,D.L., Geiger,B. and Krepler,R. (1982) *Cell*, **31**, 11–24.
Myers,M.W., Lazzarini,R.A., Lee,V.M.Y., Schlaepfer,W.W. and Nelson,D.L. (1987) *EMBO J.*, **6**, 1617–1626.
Osborn,M. and Weber,K. (1986) *Trends Biochem. Sci.*, **11**, 469–472.
Pachter,J.S. and Liem,R.K.H. (1984) *Dev. Biol.*, **103**, 200–210.
Parry,D.A.D., Steven,A.C. and Steinert,P.M. (1985) *Biochem. Biophys., Res. Commun.*, **122**, 1012–1018.
Platt,T., Files,J.G. and Weber,K. (1973) *J. Biol. Chem.*, **248**, 110–121.
Portier,M.-M., deNechaud,B. and Gros,F. (1984) *Dev. Neurosci.*, **6**, 335–344.
Pruss,R.M., Mirsky,R., Raff,M.C., Thorpe,R., Dowding,A.J. and Anderton,B.H. (1981) *Cell*, **27**, 419–428.
Ramaekers,F., Dodemont,H., Vorstenbosch,P. and Bloemendal,H. (1982) *Eur. J. Biochem.*, **128**, 503–508.
Quax,W., Van de Heuvel,R., Egberts,W.V., Quax-Jenken,Y. and Bloemendal,H. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 5970–5974.
Quinlan,R.A., Schiller,D.L., Hatzfeld,M., Achstaetter,T., Moll,R., Jorcano, J.L., Magin,T.L. and Franke,W.W. (1985) *Ann. N.Y. Acad. Sci.*, **455**, 282–306.
Shaw,G. and Weber,K. (1982) *Nature*, **298**, 277–279.
Steinert,P.M., Rice,R.H., Roop,D.R., Trus,B.L. and Steven,A.C. (1984) *Nature*, **302**, 794–800.
Steinert,P.M., Steven,A.C. and Roop,D.R. (1985) *Cell*, **42**, 411–419.

Received on February 11, 1988; revised on July 4, 1988

Note added in proof

The missing two N-terminal residues of protein B have meanwhile been obtained in collaboration with Dr K.Eckart by mass-spectroscopy of blocked N-terminal peptides. The N-terminal sequence is *N*-acetyl-threonine-serine-lysine, where the lysine corresponds in position to the lysine marked previously as 1 in Figure 3.