Characterisation and tissue-specific expression of the two keratin subfamilies of intermediate filament proteins in the cephalochordate *Branchiostoma*

Anton Karabinos[a], Dieter Riemer[a], Georgia Panopoulou[b], Hans Lehrach[b], Klaus Weber[1)a]

[a] Max Planck Institute for Biophysical Chemistry, Department of Biochemistry, Göttingen/Germany
[b] Max Planck Institute for Genetics, Berlin/Germany

Received August 31, 1999
Received in revised version October 15, 1999
Accepted October 25, 1999

Amphioxus – cephalochordates – chordates – cytoskeleton – keratins

The cloning of three intermediate filament proteins expressed at the gastrula stage (k1, Y1, X1) extends the size of the IF multigene family of *Branchiostoma* to at least 13 members. This is one of the largest protein families established for the lancelet. Sequence comparisons indicate five keratin orthologs, three of type I (E1, k1, Y1) and two of type II (E2, D1). This assignment is confirmed by the obligatory heteropolymeric polymerisation behaviour of the recombinant proteins. In line with the hetero-coiled-coil principle IF are formed by any stoichiometric mixture of type I and II keratin orthologs. In spite of the strong sequence drift chimeric IF are formed between K8, a human keratin II, and two of the lancelet type I keratins. We discuss whether the remaining 8 IF proteins reflect three additional and potentially cephalochordate-specific subfamilies. The tissue-specific expression patterns of the 5 keratins and some other IF proteins were analysed by immunofluorescence in the adult. Keratins are primarily present in ectodermally derived tissues. Developmental control of the expression of some IF proteins is observed, but three keratins (k1, Y1, D1) and an additional IF protein (X1) detected at the gastrula stage are expressed throughout the life cycle.

Abbreviations. aa Amino acid(s). – Ab Antibody(ies). – acc Accession number(s). – *B. lanceolatum/floridae, Branchiostoma lanceolatum/floridae.* – cDNA DNA complementary to RNA. – DTT Dithiothreitol. – EST Expressed sequence tag. – kDa Kilodalton(s). – IF Intermediate filament. – IFA Intermediate filament antibody. – mRNA Messenger ribonucleic acid. – PAGE Polyacrylamide-gel electrophoresis. – PCR Polymerase chain reaction. – nt Nucleotide(s). – RACE Rapid amplification of cDNA ends. – SDS Sodium dodecyl sulfate.

[1)] Prof. Dr. Klaus Weber, Max Planck Institute for Biophysical Chemistry, Department of Biochemistry, Am Fassberg 11, D-37077 Göttingen/Germany, Fax: ++55120115 78.

Introduction

The large multigene family of cytoplasmic intermediate filament (IF) proteins, which covers in man more than 50 members, is readily divided in 5 major subfamilies (for reviews see Fuchs and Weber, 1994; Parry and Steinert, 1995). Type I and type II keratins are the largest subfamilies and give rise to the epithelial keratin filaments which are based on obligatory heteromeric coiled coils formed by a type I and type II molecule. There are at least 9 type II and 12 type I keratins in the soft epithelia, and similar numbers may be expected for the type I and II keratins of hair and nails, which reflect a mammalian specialisation. Type III covers 4 mesenchymally derived proteins (desmin, vimentin, peripherin and GFAP), and 5 neuronal proteins (α-internexin, nestin and the three neurofilament proteins NF-L, NF-M and NF-H) are summarised as type IV. The two IF proteins of the eye lens (phakinin and filensin) and two large-molecular-weight proteins (synemin and paranemin) seem to fall outside these subfamilies. Finally, the type V covers the lamins, which are restricted to the nuclear compartment. All metazoan lamins so far described differ from the vertebrate cytoplasmic IF proteins by an extra 42 residues (6 heptads) present in the coil 1b subdomain of the central rod domain and by unique carboxyterminal tail domains, which display a nuclear localisation signal and in most cases a terminal CaaX motif (Stuurman et al., 1998: Erber et al., 1999).

The sequences of cytoplasmic IF proteins make an interesting contribution to metazoan molecular phylogeny. The short coil 1b subdomain extends along the chordate branch from the vertebrates to the cephalochordates and urochordates (Riemer et al., 1992, 1998; Riemer and Weber, 1998) while all cytoplasmic IF proteins of the eleven protostomic phyla analysed to date show a much closer relation to nuclear lamins (Weber et al., 1989; Dodemont et al., 1994; Erber et al., 1998). They have the longer coil 1b domain of lamins and display in most cases a lamin-like homology segment in the tail domain. Although it is not yet known whether the short coil 1b version

0171-9335/00/79/01-17 $12.00/0

is a chordate marker or a property of all deuterostomes (Erber et al., 1998), evolutionary parsimony suggests that the type I to IV divergence of the IF subfamilies was preceded by the deletion within coil 1b in the common ancestor. Thus the analysis of IF proteins from the early chordates could make important contributions as to the evolutionary roots of type I to IV subfamilies and should also provide an estimate of the complexity of the multigene family. These questions are particularly interesting in the case of *Branchiostoma* (*Amphioxus*), since cephalochordates (acrania) are generally considered to be the sister group of the vertebrates (Holland, 1996).

We previously cloned 10 distinct IF proteins from larval and adult cDNA libraries of *Branchiostoma* (Riemer et al., 1992, 1998; Karabinos et al., 1998). Here we extend the size of the multigene family to 13 by three complete clones which are based on EST sequences obtained from a gastrula library. One of them was recently also isolated from a larval library (Luke and Holland, 1999). Based on the obligatory heteropolymeric IF formation, analysed with purified recombinant proteins, we have identified two type II and three type I keratins. In spite of the sequence drift between human and lancelet proteins, filaments can be formed in vitro by a human type II keratin and two of the lancelet type I keratins. Three keratins are already expressed at the gastrula stage. We describe the tissue-specific expression patterns in the adult lancelet and discuss the possibility that the 8 additional lancelet IF proteins reflect two to three IF subfamilies which are cephalochordate specific.

Materials and methods

Branchiostoma floridae (*B. floridae*) embryos and adult animals were from Tampa, FL, USA. Adult *Branchiostoma lanceolatum* (*B. lanceolatum*) were from the island of Helgoland in the North Sea and from Roscoff, France. A larval day 2–4 *B. floridae* λ Zap II library was a gift from Dr. Linda Holland, La Jolla, CA. Other libraries were as described (Karabinos et al., 1998; Riemer et al., 1998).

Overlapping 5′and 3′cDNA sequences for E2, k1, X1 and Y1 were amplified by RACE using the Marathon cDNA amplification kit (Clontech, Heidelberg, Germany). Positions of primers in nucleotides (nt) refer to their respective cDNA sequences: E2 (accession number (acc.) AJ010293): sense nt 772–798, antisense 1321–1295; k1 (acc. AJ245426): sense 506–533, antisense 991–965; X1 (acc. AJ245427): sense 5′-CGTTGGAGGAGATGTACGCTCGCAG-3′, antisense 5′-CAACAGCTGCTTGTAGCAGGTGATCTC-3′; Y1 (acc. AJ245428): sense 5′-CCGCCGGACTTTGACGGGTCTGCAAGC-3′, antisense 5′-CTAGAGCCAGCTTGATGTTCATGAGCTCC-3′. Design of primers for X1 and Y1 amplification was based on EST sequences isolated from *B. floridae*.

Extraction of *Branchiostoma* genomic DNA and conditions for PCR amplification using 50 ng of genomic DNA as template were as described (Riemer et al., 1998). Primer sequences and locations in their respective cDNAs were as follows (for positions of X1 primers see Fig. 1): C1, sense nt 39–68, 383–410, 567–595, 843–869, antisense 410–383, 595–567, 869–843; C2, sense 164–190, 1505–1533, 1861–1889, 2065–2092, 2260–2287, 2415–2441, antisense 672–646, 1889–1861, 2092–2065, 2287–2260, 2441–2415; D1, sense 63–90, 1384–1410, 1704–1731, antisense 628–600, 1731–1704, 1935–1909; E1, sense 46–72, 639–669, 1152–1179, antisense 603–577, 669–639, 1179–1152, 1296–1273; E2, sense 79–106, 754–781, 990–1015, 1321–1348, antisense 781–754, 1015–990, 1348–1321, 1605–1578; k1, sense 74–100, 634–661, 965–991, 1150–1177, antisense 661–634, 991–965, 1177–1150, 1363–1336; Y1, sense 62–90, 636–663, 789–817, 1107–1134, antisense 663–636, 817–789, 1134–1107, 1320–1293. Introns were identified by aligning genomic and corresponding cDNA sequences. For other acc. see Riemer et al. (1998) (C1, C2, D1) and Karabinos et al. (1998) (E1).

For expression of full-length polypeptides the coding sequences of the IF cDNAs were amplified by PCR. Reaction conditions were as described (Dodemont et al., 1994). For C2 amplification the denaturing step in each cycle was extended to 30 seconds. PCR products of A3, B2, E2, k1, and Y1 were cloned into pET23 vector (Novagen, Madison, WI, USA) and of C2 into pKK388–1 (Clontech). Recombinant proteins, highly enriched in inclusion body preparations, were dissolved in 8 M urea and purified in this solvent by ion exchange chromatography (Karabinos et al., 1998). Protein purity was monitored by gel electrophoresis and automated Edman degradation. Recombinant proteins B1, E1 and D1 were as described (Karabinos et al., 1998). Aliquots of recombinant proteins (40 µl, about 0.2 mg/ml) either alone or as stoichiometric mixtures were dialysed at room temperature for 3 or more hours on dialysis filters against filament buffer containing 1 mM DTT. For D1/E1, E2/E1 and E2/k1 50 mM Tris-HCl, pH 7.6, was used. Other assembly buffers were 20 mM Tris-HCl, pH 6.8, (D1/k1), 10 mM Tris-HCl, pH 7.2 (D1/Y1), 10 mM Tris-HCl, pH 7.4 (E2/Y1), 12.5 mM Tris-HCl, pH 7 (human K8 plus Y1) and 20 mM Tris-HCl, pH 7.1, plus 10 mM NaCl (human K8 plus Y1). Negative staining with 2% uranyl acetate and electron microscopy were as described (Karabinos et al., 1998). Recombinant human keratins K5, K8, K14 and K18 were a generous gift from Dr. H. Herrmann (DKFZ, Heidelberg, Germany).

Rabbit antisera were raised with the following synthetic peptides or purified recombinant proteins as antigens: anti-E1 against the E1 peptide CGISGMWTEEKPTMRG (aa 102–116 in Fig. 2A, B); anti-E2 against the C-terminal E2 peptide CGFKGTMQSSAFGRT (Fig. 2C); anti-k1 against the C-terminal k1 peptide CTSYSTTTVTKTTKY (Fig. 2C); anti-X1 against the C-terminal X1 peptide CGLKLGKGSFL KIRR (Fig. 1); anti-Y1 against the recombinant full-length Y1 protein (see above); anti-C1 against a polypeptide covering part of the tail domain of C1 (aa 371–658 of the C1 protein in Riemer et al., 1998). Peptides were coupled via their extra N-terminal cysteine to hemocyanin. All antisera were affinity purified and if necessary also preabsorbed on other recombinant IF proteins coupled to cyanogen-bromide-activated Sepharose beads (Pharmacia, Uppsala, Sweden). The specificity of each Ab was verified by immunoblotting on a total *Branchiostoma* protein extract and on purified recombinant IF proteins as described (Riemer et al., 1998).

Frozen sections of adult *B. lanceolatum* and *B. floridae* were processed for immunofluorescence microscopy as described (Riemer et al., 1998). Antisera dilutions were: anti-C1 1:100, anti-E1 1:10, anti-E2 1:100, anti-k1 1:100, anti-Y1 1:10 and anti-X1 1:50. Murine intermediate filament Ab (IFA) was the undiluted hybridoma supernatant.

Early developmental expression of *B. floridae* IF mRNAs was determined by hybridisation using the gastrula (5–6 hours) library (Panopoulou et al., in preparation). The A2, B1, B2, C1, D1 and k1 hybridisation probes were amplified from the corresponding *B. floridae* cDNAs (Riemer et al., 1998; Luke and Holland, 1999) and cover parts of the rod domain without the terminal consensus sequences. Other hybridisation probes (E1, Y1 and X1) were amplified from both, a larval cDNA library (see above) and an adult *B. floridae* Marathon cDNA library (Clontech, Heidelberg, Germany) using primers derived from the coding sequences of the corresponding *B. lanceolatum* cDNAs. Accession number of *B. floridae* E1 is AJ245430. Grid filters of the gastrula library were hybridized overnight at 65°C in 0.5 M Na-phosphate/5% SDS buffer and washed at high stringency with 40 mM Na-phosphate/0.1% SDS buffer at 65°C for 20 min. All methods dealing with the subsequent isolation and sequencing of positive clones and PCR products were as described (Riemer et al., 1998).

Results

New IF clones

The clones MPIMGBFLG- 17O18, 115P03 and 17P02 were identified by 5′tag sequencing of clones preselected by oligonucleotide fingerprinting from a gastrula (5 to 6 h) cDNA library of *B. floridae* (Panopoulou et al. in preparation).

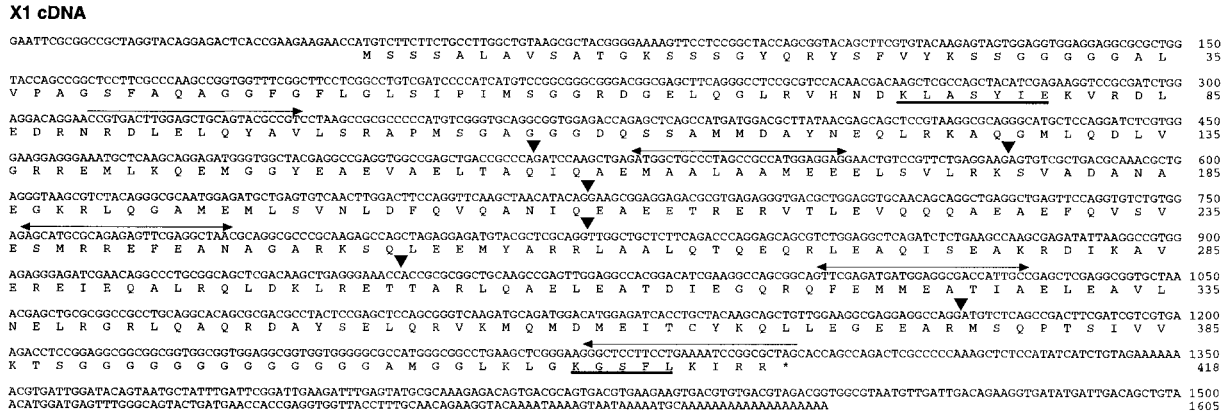


Fig. 1. Nucleotide and predicted aa sequence of *B. lanceolatum* cDNA encoding IF protein X1. The translational stop codon is marked by an *asterisk* (acc. AJ245427). *Arrows* indicate primers used for the amplification of genomic sequences. *Triangles* mark the 6 intron positions in the gene. Two peptides from spot X in the epidermis cytoskeleton (Karabinos et al., 1998) are *underlined*.

Corresponding sequences were isolated from adult *B. lanceolatum* by PCR and full-length clones were obtained by extension with RACE PCR. The three novel IF clones of *B. lanceolatum* were named X1 (acc. AJ245427), Y1 (acc. AJ245428) and k1 (acc. AJ245426). The cDNA clone encoding the rod domain of *B. lanceolatum* E2 (Karabinos et al., 1998) was used to obtain a full-length E2 clone (acc. AJ010293) by RACE extensions.

The full-length cDNA and the predicted protein sequence for X1 are shown in Fig. 1, and the predicted protein sequences for Y1, k1 and E2 are given in the sequence of alignment of Fig. 2 (see below). The k1 protein of *B. lanceolatum* corresponds to the *B. floridae* protein recently reported by Luke and Holland (1999) but clearly lacks the unusual one amino acid insertion 22 residues prior to the C-terminal end of the rod domain. Using PCR we have also generated a corresponding cDNA fragment for k1 from *B. floridae* and found that it also lacks an additional codon (acc. AJ245432). Our protein sequence in this region reads GRLKSEMS*SMLKQYEDLIK*LKLALETE instead of GRLKSEMS*QCAETVCGTSSS*LKLALETE. Thus k1 clearly conforms with the highly conserved length found for the coil 2b domain of IF proteins (Fig. 2; Fuchs and Weber, 1994; Parry and Steinert, 1995).

Sequence conservation and divergence between human and lancelet keratins

Fig. 2 gives a sequence alignment of the *Branchiostoma* keratins I (E1, k1, Y1) and II (D1, E2) with some human keratins I and II. Over the central rod domain a minor length variability is restricted to the non-helical linkers connecting coils 1a and 1b and coils 1b and 2a. These linkers also show a particularly high sequence drift between lancelet and human sequences. Keratin I and II sequences of one species show about 50 to 65% identity while between species the values are lowered to 43% (keratins II) and 34% (keratins I), respectively. Interestingly, the N-terminal half of coil 1b is much less conserved than the second half, which is particularly well conserved in type II keratins. Coil 2a and the following linker 2 show better conservation in type II than type I keratins. The consensus sequence at the C-terminal end of coil 2b, which is a marker of all IF sequences, is particularly well conserved in the last 4 to 5 heptads of type II keratins. A screen of the heptad repeat pattern (Fig. 2D) shows that the seven positions show a strikingly different degree of identical and conserved residues. Particularly poor conservation is seen for the *c* and *f* positions.

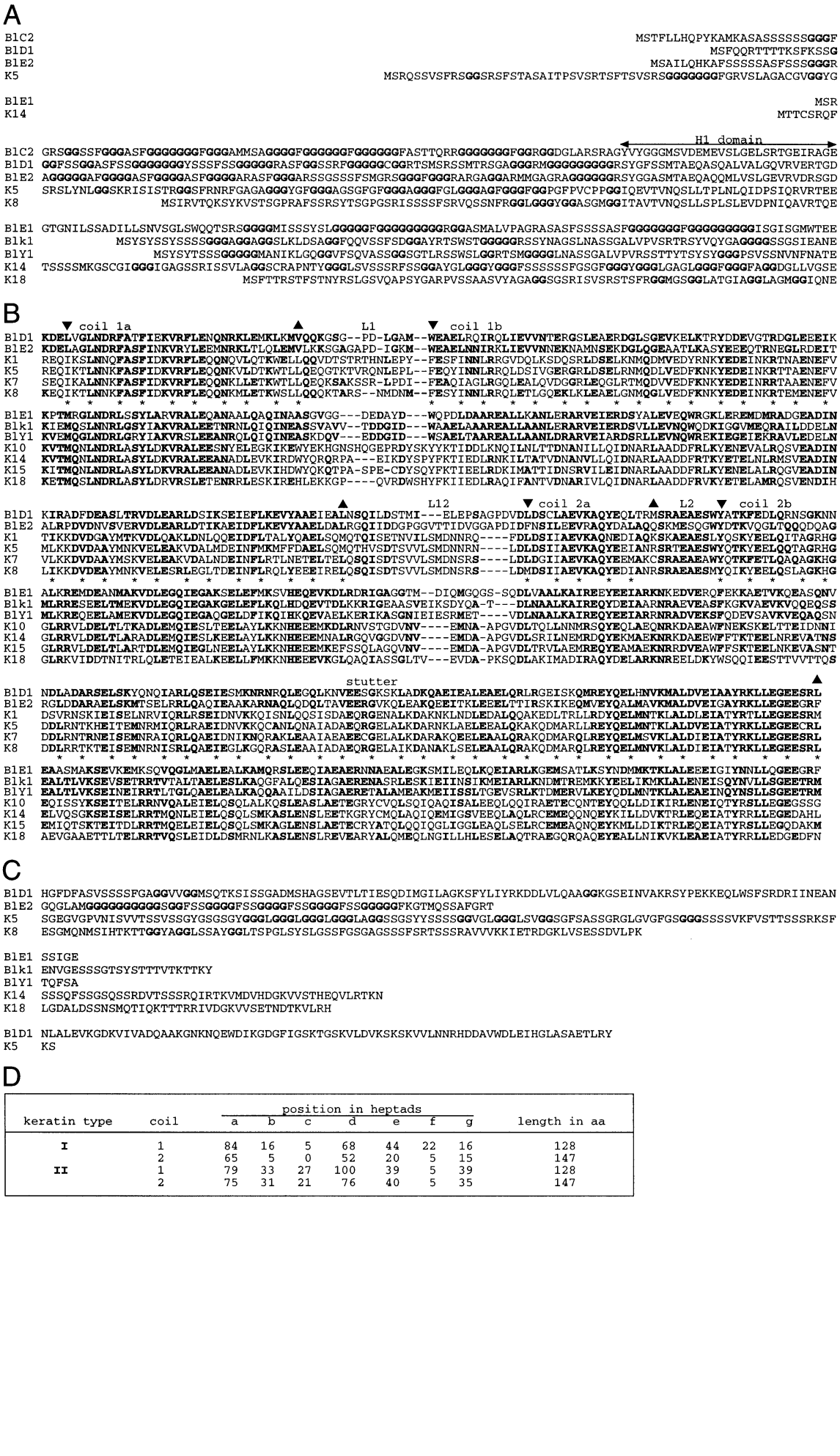
The 30 residues of the non helical H1 domain, located in front of the rod domain in human type II keratins, are highly conserved and somehow involved in filament assembly, since point mutations in H1 lead to hereditary human epidermal fragility syndromes (Parry and Steinert, 1995). Interestingly, the H1 region of the lancelet type II keratins shows a strong divergence from the human sequences (Fig. 2). In type I keratins of man and *Branchiostoma* the corresponding region is rich in glycine residues.

Mammalian type I and II keratins from stratified squamous epithelia, but not from simple epithelia, show pronounced glycine loops in the terminal domains. This sequence motif, defined as a series of glycine and serines flanked by aromatic and/or large apolar residues, is thought to provide a conformation of great flexibility (Parry and Steinert, 1995). Although the lancelet seems to contain only simple epithelia (Bartnik and Weber, 1989; Ruppert, 1997) most of the sequence of the tail domain of the epidermal protein E2 is provided by 5 consecutive glycine loops (Fig. 2). D1, the second lancelet type II keratin, has an unusually large tail domain (184 residues) with unique sequences. Of the three type I keratins, E1 and Y1 are essentially tail-less (5 residues) like human keratin K19 (13 residues) while lancelet k1 has a mini-tail domain (23 residues).

The N-terminal head domains of lancelet keratins differ in length but except for their first region they are all very rich in glycine residues. These are sometimes arranged in glycine loops with flanking large apolar residues (see above) as in D1 but more often as glycine strings. These regions harbour, in addition to glycine and serine, occasional threonine, alanine and arginine residues and often have only one large apolar amino acid as flanking residue. Curiously, both glycine loops and strings also occur outside the keratins in the terminal domains of lancelet proteins B1, C2 and X1.

Branchiostoma keratins identified by obligatory heteropolymer filament formation

Mammalian keratin filaments are obligatory heteropolymeric filaments based on equal numbers of keratin I and II molecules arranged as heteromeric double-stranded coiled coils (Hatzfeld and Weber, 1990; Parry and Steinert, 1995). We have made


A
B
C
D
H1 domain
coil 1a
L1
coil 1b
L12
coil 2a
L2
coil 2b
stutter
keratin type
coil
position in heptads
a b c d e f g
length in aa
I 1 84 16 5 68 44 22 16 128
2 65 5 0 52 20 5 15 147
II 1 79 33 27 100 39 5 39 128
2 75 31 21 76 40 5 35 147

◀ **Fig. 2.** Alignment of *Branchiostoma* type I and II keratins with human keratins. The three-domain structure (head, rod, tail; A, B, C) is indicated. Above the sequence blocks of the rod domains of type II keratins the ends of the subdomains (coils 1a, 1b, 2a and 2b) are marked by *arrowheads* and the non-helical linkers are indicated (B). *Asterisks* below the sequence blocks of type II keratins mark the *a* and *d* positions of the consecutive heptads (*a* to *g*) with a heptad stutter in coil 2b (for details on IF structure see Fuchs and Weber, 1994; Perry and Steinert, 1995). *Dashes* are used for optimisation of the sequence alignments and to cover the structural relation between type I and II keratins. Type II keratins are the *Branchiostoma* (B1) proteins D1 and E2 and human keratins 1, 5, 7, and 8. Type I keratins are *Branchiostoma* (B1) proteins E1, k1 and Y1 and human keratins 10, 14, 15, and 18. Note that length variability is restricted to linkers L1 and L12. *Bold letter type* is used for identical residues in the *Branchiostoma* and human rod sequences. Over the head and tail domains the number of type I and II human keratins is reduced from 4 to 2. Note the H1 domain (*arrow*) prior to the rod domain of type II keratins and the strong sequence drift between human and *Branchiostoma* sequences along this domain. Note the similarity in the H1 domain of the lancelet keratins II and the C2 proteins. Glycine-rich regions in the terminal domains (A, C) are indicated in *bold* type. Part D gives the percentages of identical plus conserved amino acid residues along the heptad repeats for human plus lancelet keratins.

use of this molecular feature of keratins (Fig. 3) parallel to the sequence data (Fig. 2). IF proteins D1, E1, E2, k1, and Y1 were expressed in *E. coli*, and the recombinant proteins present in isolated inclusion bodies were dissolved in 8 or 9 M urea and purified in this solvent by ion exchange chromatography on Mono Q and Mono S columns. Homogenous proteins were dialysed either alone or in stoichiometric mixtures to remove the urea and the self-assembly products were monitored by electron microscopy after negative staining. While each individual protein yielded only aggregated material, the following mixtures provided the IF shown in Fig. 3: E1/E2, k1/E2, Y1/E2, k1/D1 and Y1/D1. Taken together with the previous reconstitution of E1/D1 filaments (Karabinos et al., 1998) each type II protein (E2, D2) forms IF when mixed with any type I protein (E1, k1, Y1).

We also explored the possibility of obtaining chimeric IF from human and *Branchiostoma* keratins. Thus the type II human K8 and *Branchiostoma* E1, a type I keratin, resulted in relatively good looking but not perfect IF (Fig. 3) although E1 and human K18, the natural partner of K8, share only 33% sequence identity over their rod domains (see Fig. 2). Mixtures of human K8 plus *Branchiostoma* Y1 also lead to IF (Fig. 3) while human K18 plus *Branchiostoma* D1 interactions seem already arrested at the tetrameric level revealed in metal-shadowed specimens of glycerol sprayed material (data not shown). These results show, even without looking at the sequences, that *Branchiostoma* proteins E1/Y1 and D1 behave in IF assembly as keratins I and II, respectively. We note, however, that several other mixtures failed to yield filaments at least under the limited set of conditions tried. These include human K8 or K5 with *Branchiostoma* k1 and human K18 or K14 with *Branchiostoma* E2 and D1. Human K5 failed also to form filaments with *Branchiostoma* E1 and Y1 (see Discussion).

Evolutionary tree

Figure 4 shows an evolutionary tree generated from the rod sequences of all 13 *Branchiostoma* IF proteins and a variety of human representatives for vertebrate type I to IV proteins.

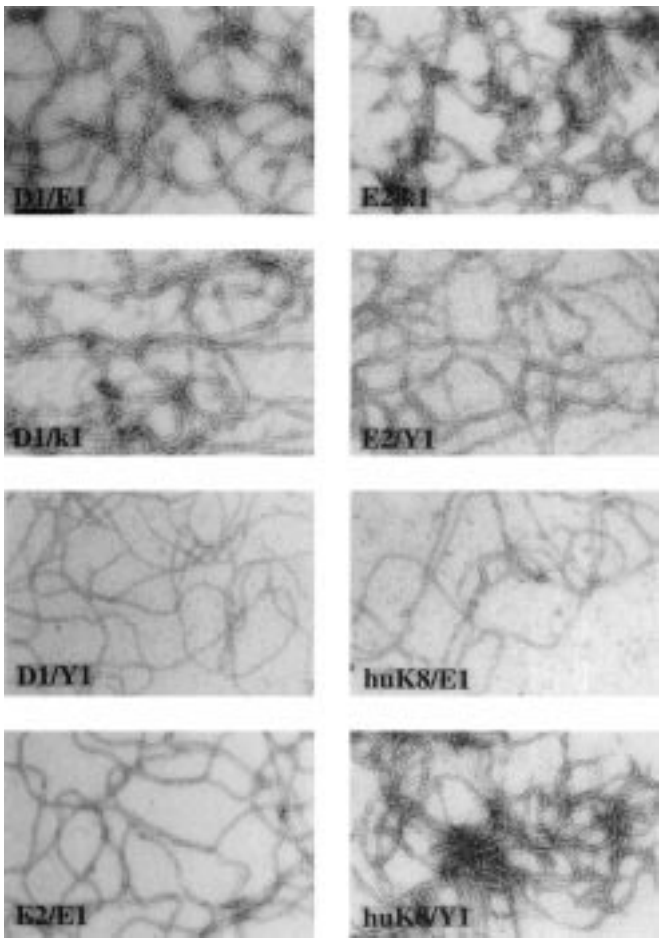


Fig. 3. In vitro filament formation by recombinant *Branchiostoma* keratins. *Branchiostoma* type I keratins (E1, k1, Y1) mixed with an equimolar amount of a *Branchiostoma* type II keratin (D1, E2) form IF upon removal of urea. Electron micrographs were taken after negative staining with uranyl acetate (bar in top left corner is 0.2 μm). Individual proteins do not form IF. Note that the human type II keratin 8 (hu K8) forms IF with the lancelet type I keratins E1 and Y1 (hu K8/E1; hu K8/Y1).

Included is also the muscle IF protein established for the urochordate (tunicate) *Styela plicata* (Riemer and Weber, 1998). This tree differs from previous versions (Karabinos et al., 1998; Luke and Holland, 1999) by the availability of a much larger group of lancelet IF proteins, by the inclusion of a true type III protein from an early chordate, i.e. the tunicate *Styela*, and by the identification of the lancelet X1 protein as the most diverse chordate IF protein. Of the 13 lancelet proteins, 5 are unambiguously identified in their relation to vertebrate IF subfamilies. In agreement with the detailed sequence features (Fig. 2) and the obligatory heteropolymeric IF assembly data (Fig. 3) proteins E1, k1 and Y1 of the lancelet form a sister group with human keratins I, while proteins D1 and E2 form the sister group of human keratins II. Independent support for this view comes from the observation of specific assembly products in mixtures of lancelet and human keratins (see above).

The five *Branchiostoma* IF proteins of the A/B group (Riemer et al., 1998) appear as a separate branch. Similarly, the two proteins C1 and C2 define a separate branch of lancelet IF proteins and also share unusual features in their tail

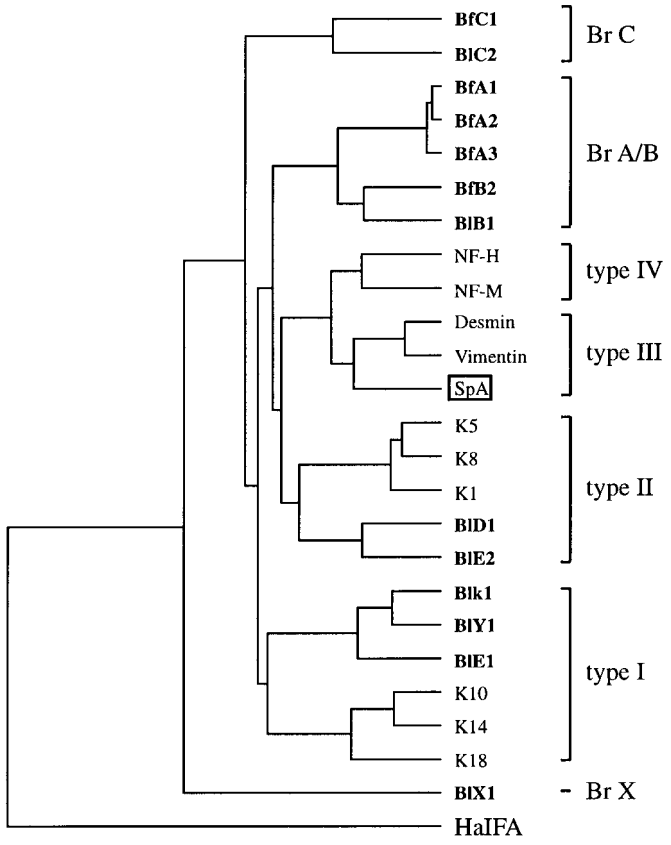


Fig. 4. Reconstruction of an evolutionary tree formed by the sequences of the rod domains of 13 IF proteins from the cephalochordate *Branchiostoma* and human type I to IV proteins. The human proteins serving as representatives for vertebrate proteins are from the EMBL/GenBank and their acc. have been summarised (Karabinos et al., 1998). Except for A1, A2, A3, B2, and C1, which are from *B. floridae* (Bf), all other lancelet sequences are from *B. lanceolatum* (Bl). Except for k1, Y1, E2, and X1, which were determined in this study, all other *Branchiostoma* proteins as well as their acc. have been described (Riemer et al., 1998; Karabinos et al., 1998). The IF protein of the snail *Helix pomatia* (HaIFA; P16275) represents the protostomic IF proteins (see Introduction) and served as outgroup. The four subfamilies I to IV of vertebrate IF proteins are indicated at the right. Note the sister group relationship between *Branchiostoma* proteins E1, k1, Y1 and human keratins 10, 14, 18 marked as type I keratins and the corresponding relationship between *Branchiostoma* proteins D1, E2 and human keratins 1, 5 and 8 marked as type II keratins. Two additional branches of *Branchiostoma* proteins are indicated as the Br A/B branch and the Br C branch respectively (Karabinos et al., 1998). The *Branchiostoma* X1 protein shows the most divergent sequence among the chordate proteins. Note that the muscle IF protein SpA (*boxed*) from the urochordate *Styela plicata* (AJ005020; Riemer and Weber, 1998) falls into the type III group. Note also the sister group relation between type III and type IV proteins.

domains (Riemer et al., 1998). Finally, the *Branchiostoma* protein X1 (Fig. 1) has the most remote sequence among the proteins used in constructing the tree given in Fig. 4.

Organisation of IF genes

We previously characterised the exon/intron patterns in the complete B1 gene (Riemer et al., 1992) and in the head- and rod-encoding regions of the A1, A2, B2, D1, and C2 genes (Riemer et al., 1998). Figure 5 extends this analysis with the complete genes E1, E2, k1, Y1, and X1 and provides also the tail-encoding regions of D1 and C2 as well as the rod-encoding region of C1. All 12 *Branchiostoma* genes show over the rod domain intron patterns similar to that of vertebrate type III genes, since even the *Branchiostom*a keratin I (E1, k1, Y1) and keratin II (D1, E2) genes lack the intron positions which are type specific in vertebrate type I and II genes. A few unusual intron positions are marked in Fig. 5. Of particular interest is the sixth intron of the E2 gene which is not situated at the end of the rod domain but moved by 27 nucleotides and thus occupies a position previously thought to be highly specific for neuronal type IV genes (Lewis and Cowan, 1986). The large tail domains of the D1 and C2 genes are rich in introns. None of the *Branchiostoma* genes resembles in organisation the vertebrate type IV genes which have only very few introns (Fig. 5).

Tissue-specific expression of cytoplasmic IF proteins

To characterise the expression patterns of the *Branchiostoma* keratins E1, E2, k1, and Y1 (for D1 see Riemer et al., 1998) and the non-keratin IF proteins C1 and X1 (for B1 and C2 see Riemer et al., 1998) antibodies were raised in rabbits. The specificity of these purified antibodies for a single IF protein is demonstrated in Fig. 6. Antibodies to E1, E2 and k1 recognised in immunoblots the recombinant proteins and corresponding comigrating polypeptides in the total protein extract of the adult lancelet. The Y1 antibody detected the recombinant protein. The lack of a reaction on the total protein extract is most likely due to the small amount of Y1 (see below). Antibodies to C1 and X1 recognised single polypeptides of approximately 120 kDa (C1) and 50 kDa (X1) in total protein extracts in good agreement with the molecular masses calculated from the C1 and X1 sequences (Riemer et al., 1998).

Immunofluorescence microscopy on frozen sections of adult *Branchiostoma* (Fig. 7) showed coexpression of E1, E2 and X1 in the epidermis in agreement with the two dimensional gel patterns of cytoskeletal preparations from dissected epidermis, which also revealed the presence of D1 and C2 (Karabinos et al., 1998). E1, E2 and X1 are also present in cutaneous canals and in the nerve cord. In the nerve cord the staining patterns are not identical indicating that the three proteins do not colocalise. E2 and X1 reactivities are distributed over the central grey and peripheral white matter as previously seen for C2 and D1 (Riemer et al., 1998). E1 fluorescence coincided with the perikarya of cells surrounding the central canal and extended into the periphery as thin projections. Since no reliable differential markers for neurones and glial cells are available for *Branchiostoma* (see Ruppert (1997) for lancelet neurohistology), we cannot discriminate between E1 expression in neuronal and glial cells.

C1 and k1 staining occurred exclusively in atrial epithelia. Other ectodermal tissues such as epidermis and nerve cord showed no reaction. Y1 expression was most prominent in cutaneous canals and significantly reduced in the nerve cord. No Y1 staining was observed in the epidermis. Several mesodermally derived tissues (musculature, notochord and coelomic epithelia) and tissues of endodermal origin (inner part of gill slits, intestine and gonads) were negative with all antibodies used in this study.

Expression of IF genes at the gastrula stage

Using 5′tag sequencing of clones preselected by oligonucleotide fingerprinting from a gastrula cDNA library of *B. floridae*

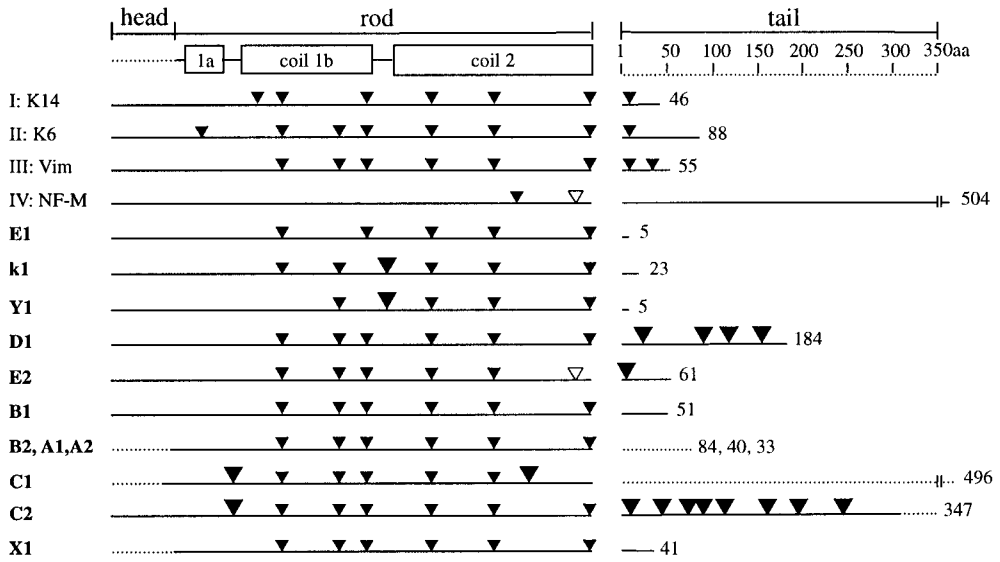


Fig. 5. Comparison of intron positions in 12 IF genes of *Branchiostoma* and vertebrate IF type I to IV genes. The tripartite structural organisation of the cytoplasmic IF proteins is indicated at the top. Numbers along the tail domain give the size in aa. Dots indicate parts of the genes not yet analysed. *Small and large filled arrowheads* specify common and unique intron positions, respectively. An *open arrowhead* is used to highlight the same position of the second intron in the vertebrate neurofilament gene NF-M and the sixth intron in the *Branchiostoma* E2 gene. In the following, intron positions are listed by the nt numbers of the cDNA sequences after which the introns are inserted. These are for E1 620, 777, 945, 1071, and 1292; for k1 621, 682, 837, 943, 1069, and 1290; for Y1 675, 830, 936, 1062, and 1283; for E2 738, 799, 895, 1072, 1198, 1392, and 1449; for C1 209, 364, 425, 521, 692, 806, and 916; for the tail regions of D1 1477, 1652, 1785, and 1884; for the tail region of C2 1574, 1643, 1714, 1775, 1865, 2005, 2104, and 2253. For X1 see Fig. 1. Except for the C1 gene from *B. floridae* all other genes are from *B. lanceolatum*. For all other intron positions see Riemer et al., 1998.

led to EST sequences for the two new type I keratins k1 and Y1 (see above), the already known type II keratin D1 and the IF protein X1. Using cDNA probes in hybridisation, the following IF cDNAs were not detected in the gastrula library: A2, B1, B2, C1, and E1. Thus, the gastrula expresses at least keratins k1, Y1 and D1 as well as IF protein X1.

Discussion

Based on three EST sequences from a gastrula cDNA library of *B. floridae* and a partial sequence for an epidermal IF protein of adult *B. lanceolatum* we have extended the number of complete cDNAs for cytoplasmic IF proteins from 9 (Riemer et al., 1998; Karabinos et al., 1998) to 13. This makes the IF group one of the largest multigene families so far established for the lancelet. Thirteen IF genes is still a minimal estimate and we expect that the family will still grow by a few new members and thus reach a value corresponding to one third of the human family, which has around 50 members (see Introduction).

Starting with the recent indications for *Branchiostoma* keratins (Karabinos et al., 1998; Luke and Holland, 1999) we have made a systematic analysis to identify type I and II keratins by detailed sequence comparisons and the assembly properties of recombinant proteins. The obligatory heteropolymeric keratin filaments of mammals are based on a heteromeric coiled coil formed by one type I and one type II keratin polypeptide (Hatzfeld and Weber, 1990; Parry and Steinert, 1995). Thus in vitro filament assembly is a highly sensitive assay to directly identify keratin orthologs. Sequence comparisons and the calculated evolutionary tree (Figs. 2, 4) indicate two groups (E1/k1/Y1 and D1/E2) in which individual members share 52 to 65% identity over the rod domain. The electron micrographs in Fig. 3 document that, just as with mammalian keratins, any member of the first group forms IF when mixed in stoichiometric amounts with any member of the second group, while the single proteins fail to form filaments. Since the human type II keratin K8 and the lancelet proteins E1 and Y1 can also form IF in vitro, the E1/k1/Y1 group is directly identified as type I keratins in full agreement with the interpretation of the sequences and the calculated evolutionary tree. Although we have not obtained IF with human type I keratin K18 and a *Branchiostoma* type II keratin (D1, E2) the assembly via a heteromeric coiled coil proceeds at least to the tetramer level. We also note that human K5 and the three *Branchiostoma* type I keratins as well as human K14 and the two type II lancelet keratins yielded so far no filaments.

Over the rod domains, type I and II keratins of *Branchiostoma* and man share only 35 and 44% sequence identity, and Fig. 2 identifies the regions with the strongest sequence drift. We expect that this drift will be understood, once an atomic model of an IF protein becomes available in the future. Particularly striking is also the sequence drift between the H1 domains of human and lancelet type II keratins (Fig. 2). This domain, situated just prior to the rod domain, is directly involved in keratin filament assembly (Parry and Steinert, 1995). Currently, we do not know which sequence differences are responsible for the observation that not all mixtures of *Branchiostoma* and human keratins can proceed to the level of filaments. Thus, human K18 plus lancelet D1 formed only tetramers while human K8 plus lancelet E1 or Y1 yielded long filaments.

In contrast to the various epithelial organisation of vertebrates all *Branchiostoma* epithelia including the epidermis are

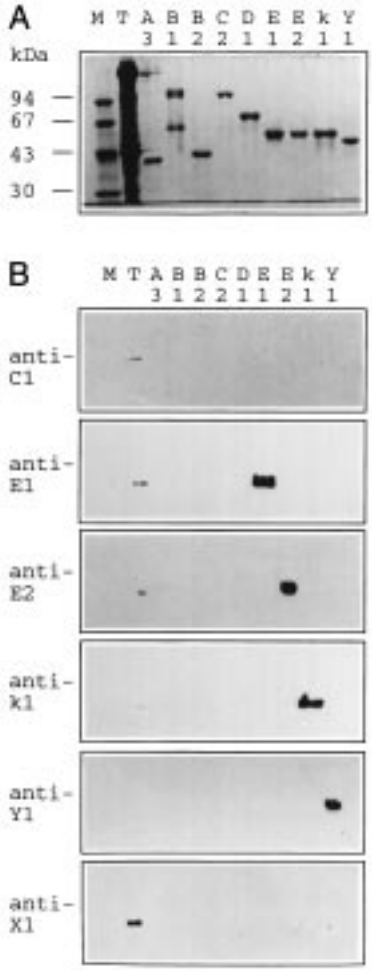


Fig. 6. Immunoblot analyses of total *Branchiostoma* extract and recombinant proteins. Equal amounts of total (T) and recombinant IF proteins (A3, B1, B2, C2, D1, E1, E2, k1, Y1) were separated by SDS-PAGE (10%) and stained with Coomassie (**A**) or transferred to nitrocellulose membranes (**B**). Blots were incubated with affinity-purified rabbit Ab to *Branchiostoma* C1, E1, E2, k1, Y1, and X1 proteins. The E1, E2 and k1 Ab recognised the corresponding recombinant protein and a single band of similar molecular mass in total *Branchiostoma* extract (note a very weak k1 signal in the total extract) and did not crossreact with the other recombinant *Branchiostoma* proteins. Antibody to Y1 detected only recombinant Y1 without a reaction on the total extract. The specific Ab for C1 and X1 recognised a single band of the expected size for C1 and X1 proteins in the total extract. These two proteins are not available in recombinant form. Marker proteins (M) and an approximate molecular mass standard in kDa are given at the left. All experiments are on *B. lanceolatum* except for C1 where *B. floridae* was used.

simple epithelia (Bartnik and Weber, 1989; Ruppert, 1997). In mammals type I and II keratins from stratified squamous epithelia differ from the keratins of simple epithelia by pronounced glycine loops present in one or both terminal domains. This sequence motif, defined as a row of glycines and serines flanked by aromatic and large apolar residues, provides a conformation of great flexibility which is thought to provide a protective barrier and to stabilise the epithelium against mechanical stress (Parry and Steinert, 1995). Interestingly, the epidermal keratin E2 of the lancelet shows 5 glycine loops in the tail domain while D1, the other type II keratin, has a long and unique tail domain. Extended glycine-rich sequences with a less strict sequence requirement than the glycine loops are found in the N-terminal head domains of all lancelet keratins. Curiously, the 3 *Branchiostoma* type I keratins either lack the C-terminal tail domain (E1, Y1), as does human keratin K19, or the domain is rather short (k1).

The evolutionary tree (Fig. 4) identifies two additional branches of *Branchiostoma* IF proteins, which in contrast to the type I and type II keratins are currently difficult to relate to the vertebrate type I to IV families. The A/B branch covers A1, A2, A3, B1, and B2 while the C branch combines C1 and C2. Of these two branches the C proteins may indeed be a cephalochordate-specific lineage, since they also have unique large tail domains with α-helical sequences, which indicate coiled coil forming ability (Riemer et al., 1998). We expect that these proteins will be better understood, once immunoelectron microscopy can be used to decide for instance whether C2 is contained in the keratin filaments of the epidermis or involved in separate structures. Interestingly, inspection of the sequences shows that C1 and particularly C2 have H1 domains highly related to those of *Branchiostoma* type II keratins (Fig. 2).

Based on the ability of the lancelet B1 protein to form homopolymeric IF in vitro we have previously speculated that the A/B branch could reflect type III orthologs and that their separation versus the vertebrate type III proteins in the evolutionary tree could be due to a particularly strong evolutionary drift for this branch (Karabinos et al., 1998). However, the urochordate *Styela* has a smooth muscle IF protein which is well related to vertebrate type III proteins (Riemer and Weber, 1998) and clearly incorporated into the tree as a sister group of vertebrate type III proteins (Fig. 4). This raises the question as to a similar, not yet detected, IF protein in the lancelet. Additional information may come from future studies on the cell- and tissue-specific expression patterns for the 5 proteins of the A/B group and the use of immunoelectron microscopy. IF protein X1 poses a special problem, since it shows the most remote sequence among all chordate proteins (Fig. 4). This is particularly interesting since X1 is already expressed at the gastrula stage. In this respect we note that the gastrula of the lancelet expresses at least 4 IF proteins (X1, the two type I keratins k1 and Y1 and the type II keratin D1) while the mammalian gastrula is thought to show only keratins K8 and K18 (Jackson et al., 1980; Baribault and Oshima, 1991; Moll et al., 1982).

The current catalogue of 13 lancelet IF proteins lacks obvious orthologs for vertebrate type IV proteins. Although their rod domains are relatively close to type III proteins (Fig. 4), at least the neurofilament triplet proteins have extended acidic tail domains rich in glutamic acid and lysine, and the type IV genes have an unusually low number of introns. In contrast all IF genes of the lancelet are rich in introns (Fig. 5). To account for the low number of introns in neuronal type IV genes it is thought that the first type IV gene arose by a mRNA transposition event which removed the old introns and allowed the subsequent acquisition of a few introns at novel sites (Lewis and Cowan, 1986). While this hypothesis may still hold, we note that intron 6 of the lancelet gene E2, which encodes a type II keratin, occupies precisely a position previously thought to be specific for type IV genes. It remains to be elucidated whether neurofilament type IV proteins are an acquisition of the vertebrate lineage just like IF subfamilies A/B and C seem to be cephalochordate specific.

Using immunofluorescence microscopy we previously analysed the expression pattens of the IF proteins B1, D1 and C2

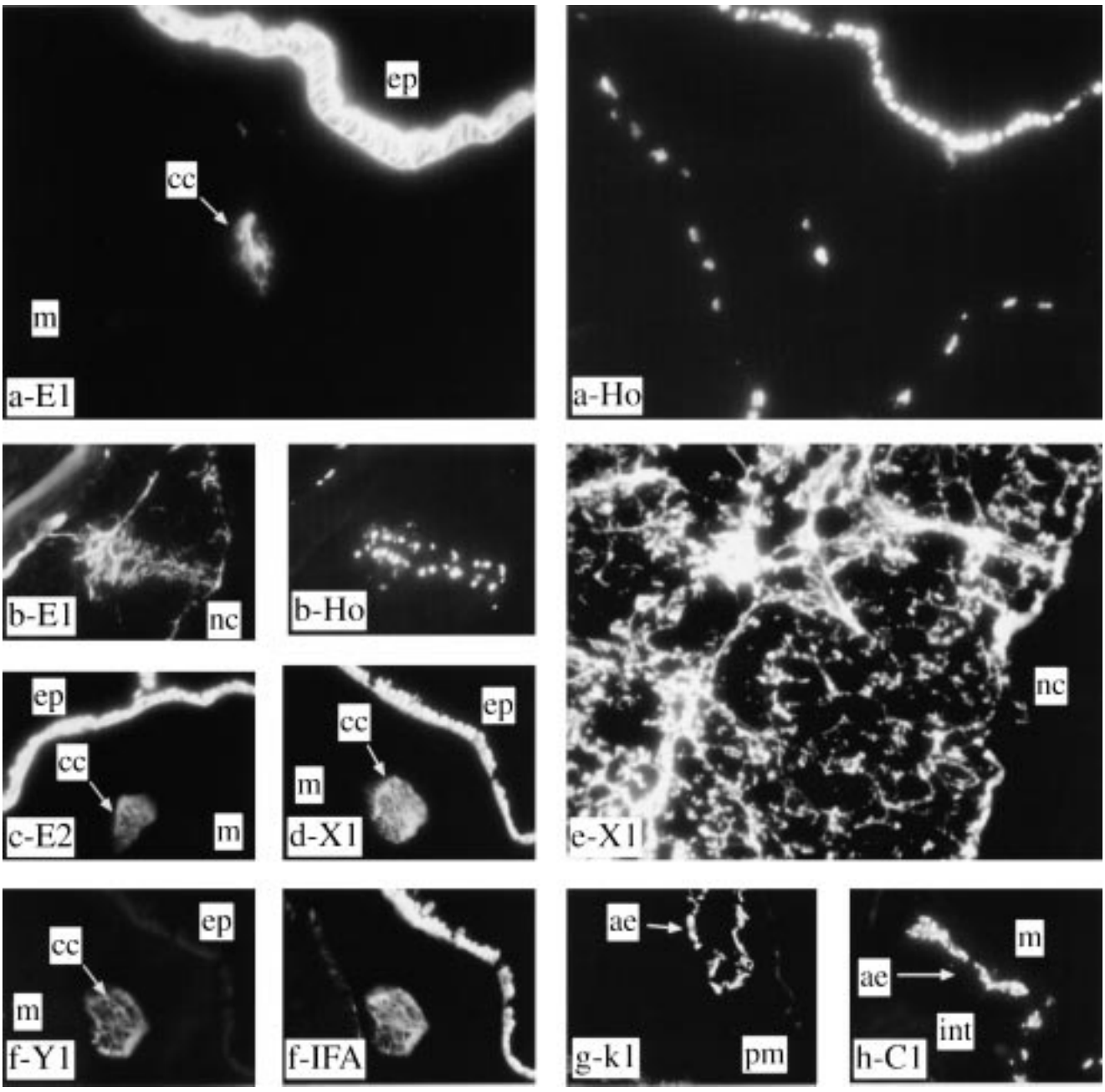


Fig. 7. Tissue-specific expression of IF proteins in adult *Branchiostoma*. Frozen sections were double-labeled with affinity-purified rabbit Abs specific for keratins E1, E2, k1, Y1, or IF proteins C1 and X1 (see Fig. 6) and with murine IFA. Staining of nuclei in the same sections was performed with Hoechst 33258 (Ho) to facilitate identification of cells and tissues. **a-E1**, **c-E2**, **d-X1.** Intensive staining of E1, E2 and X1 was observed in the epidermis (ep) and in the cutaneous canals (cc) but not in musculature (m). **a-Ho.** Hoechst staining of DNA in the epidermis, cutaneous canal and in the epithelia lining the musculature. **b-E1.** Cross section of the neural cord (nc) shows E1 expression in a subset of cells. E1 reactivity colocalised partially with cells lining the central canal (**b-Ho**). **e-X1.** Punctate staining pattern of X1 in the neural cord. **f-Y1.** Y1 immunoreactivity was found in cutaneous canals but not in epidermis (ep) or musculature (m). Y1 decoration did not completely coincide with the IFA staining pattern in these tissues (**f-IFA**). **g-k1.** Expression of k1 was identified exclusively in the atrial epithelium (ae) (pm, pterygial muscle). The same staining pattern was observed for C1 (**h-C1**). Note the absence of C1 reactivity in intestine (int) and musculature (m). C1 staining was performed on *B. floridae.* All other cross sections were from *B. lanceolatum*. Magnifications: a-E1, a-Ho, e-X1 × 2000; others × 400.

in the adult lancelet (Riemer et al., 1998). This study adds the results on six further IF proteins (E1, E2, k1, X1, Y1, C1). Thus except for A1, A2, A3, and B2 the cell- and tissue-specific expression patterns are known for the various *Branchiostoma* IF proteins. The combined results show the coexpression of keratins D1, E1, E2 together with IF proteins X1 and C2 in cells and tissues of ectodermal origin such as the epidermis and the dorsal nerve cord, while the atrial epithelium, which is also ectodermally derived shows coexpression of keratins k1 and D1 together with IF proteins C1 and C2. The immunological characterisation of the IF complement of the epidermis covering E1, E2, D1, and proteins C2 and X1 is in full

agreement with the earlier results based on two dimensional gels (Karabinos et al., 1998). The so far negative immunological results on the intestinal epithelium raises the question of additional not yet discovered keratins of the lancelet.

Tissues of mesodermal origin are generally not decorated by keratin antibodies in the lancelet. Thus the mesodermal epithelia (the epithelial lining of the fin box coelom and the myomeres) stain positively for B1, which is a non-keratin IF protein (see above and Riemer et al., 1998). However, the epithelial cells of the cutaneous canals contain keratins E1, E2 and Y1 together with proteins X1 and C2. Although some embryological data support the hypothesis that the canal epithelium is a derivative of the mesoderm, a definitive answer for its derivation is not available (Ruppert, 1997).

In the immunological survey of *Branchiostoma* tissues we found the keratins D1, E1, E2, and Y1 together with the unassigned IF proteins B1, C2 and X1 in the nerve cord. The cellular expression patterns of these 7 IF proteins seem to differ significantly. While B1 (see Riemer et al., 1998) and E1 staining was predominantly associated with the ependymal cells lining the central canal, IF proteins C2, D1, E2, and X1 relate to the thick filaments in the axonal processes which constitute the white matter of the nerve cord. The detection of true keratins (D1, E1, E2) in the nerve cord of adult *Branchiostoma* is interesting from a phylogenetic point of view. Keratins have been found in the spinal cord and certain brain regions of the shark (Schaffeld et al., 1998), and the glial cells of teleost fishes are rich in keratins, while higher vertebrates lack this keratin expression (Markl and Schechter, 1998). Therefore, keratin expression in the nerve tissue of cephalochordate and lower vertebrates may represent a phylogenetically original situation where keratins are ubiquitously found in cells of ectodermal origin. An unexpected result of the expression patterns concerns the proteins C1, C2 and X1 which have been discussed above as potential cephalochordate-specific branches of IF proteins. Their presence together with bona fide keratins in various ectodermally derived tissues raises the question of whether they can incorporate into keratin filaments or form separate structures. This question can be answered in the future by immuno electronmicroscopy.

Acknowledgements. We thank Jürgen Schünemann, Tomma Eisbein and Wolfgang Berning-Koch for expert technical assistance. Dr. Linda Holland kindly provided a larval 2–4 day *B. floridae* λ Zap II library. Recombinant human keratins K5, K8, K14, and K18 were kindly provided by Dr. H. Herrmann, DKFZ, Heidelberg, Germany.

References

Baribault, H., Oshima, R. G. (1991): Polarized and functional epithelia can form after the targeted inactivations of both mouse keratin 8 alleles. J. Cell Biol. **115**, 1675–1684.

Bartnik, E., Weber, K. (1989): Widespread occurrence of intermediate filaments in invertebrates: common principles and aspects of diversification. Eur. J. Cell Biol. **50**, 17–33.

Dodemont, H., Riemer, D., Ledger, N., Weber, K. (1994): Eight genes and alternative RNA processing pathways generate an unexpectedly large diversity of cytoplasmic intermediate filament proteins in the nematode *Caenorhabditis elegans*. EMBO J. **13**, 2625–2638.

Erber, A., Riemer, D., Bovenschulte, M., Weber, K. (1998): Molecular phylogeny of metazoan intermediate filament proteins. J. Mol. Evol. **47**, 751–762.

Erber, A., Riemer, D., Hofemeister, H., Bovenschulte, M., Stick, R., Panopoulou, G., Lehrach, H., Weber, K. (1999): Characterisation of the *Hydra* lamin and its gene; a molecular phylogeny of metazoan lamins. J. Mol. Evol. **49**, 260–271.

Fuchs, E., Weber, K. (1994): Intermediate filaments: Structure, dynamics, function and disease. Annu. Rev. Biochem. **63**, 345–382.

Hatzfeld, M., Weber, K. (1990): The coiled coil of in vitro assembled keratin filaments is a heterodimer of type I and II keratins: use of site-specific mutagenesis and recombinant protein expression. J. Cell Biol. **110**, 1199–1210.

Holland, P. W. H. (1996): Molecular biology of lancelets: insights into development and evolution. Israel J. Zool. **42**, 247–272.

Jackson, B. W., Grund, C., Schmid, E., Bürki, K.; Franke, W. W., Illmensee, K. (1980): Formation of cytoskeletal elements during mouse embryogenesis. Differentiation **17**, 161–179.

Karabinos, A., Riemer, D., Erber, A., Weber, K. (1998): Homologues of vertebrate type I, II and III intermediate filament (IF) proteins in an invertebrate: the IF multigene family of the cephalochordate *Branchiostoma*. FEBS Lett. **437**, 15–18.

Lewis, S. A., Cowan, N. J. (1986): Anomalous placement of introns in a member of the intermediate filament multigene family: an evolutionary conundrum. Mol. Cell Biol. **6**, 1529–1534.

Luke, G. N., Holland, P. W. H. (1999): An amphioxus type I keratin cDNA and the evolution of intermediate filament genes. J. Exp. Zool. **285**, 50–56.

Markl, J., Schechter, N. (1998): Fish intermediate filament proteins in structure, evolution and function. In: Hermann, H. and Harris, J. E. (Eds.): Subcellular Biochemistry Vol. 31. Plenum Press, New York, pp. 1–33.

Moll, R., Moll, I., West, W. (1982): Changes in the pattern of cytokeratin polypeptides in epidermis and hair follicles during skin development in human tissues. Differentiation **23**, 170–178.

Parry, D. A. D., Steinert, P. M. (1995): Intermediate filament structure. Springer, New York.

Riemer, D., Dodemont, H., Weber, K. (1992): Analysis of the cDNA gene encoding a cytoplasmic intermediate filament (IF) protein from the cephalochordate *Branchiostoma lanceolatum*; implications for the evolution of the IF protein family. Eur. J. Cell Biol. **58**, 128–135.

Riemer, D., Karabinos, A., Weber, K. (1998): Analysis of eight cDNAs and six genes for intermediate filament proteins in the cephalochordate *Branchiostoma* reveals differences in the multigene families of lower chordates and the vertebrates. Gene **211**, 361–373.

Riemer, D., Weber, K. (1998): Common and variant properties of intermediate filament proteins from lower chordates and vertebrates; two proteins from the tunicate *Styela* and the identification of a type III homologue. J. Cell Sci. **111**, 2967–2975.

Ruppert, E. E. (1997): Cephalochordata (Arancia), Microscopic Anatomy of Invertebrates Vol. 15, Wiley-Liss, New York, pp. 349–504.

Schaffeld, M., Löbbecke, A., Lieb, B., Markl, J. (1998): Tracing keratin evolution: catalog, expression pattern and primary structure of shark (*Scyliorhinus stellaris*) keratins. Eur. J. Cell Biol. **77**, 69–80.

Stuurman, N., Heins, S., Aebi, U. (1998): Nuclear lamins: their structure, assembly and interactions. J. Struct. Biol. **122**, 42–66.

Weber, K., Plessmann, U., Ulrich, W. (1989): Cytoplasmic intermediate filament proteins of invertebrates are closer to nuclear lamins than are vertebrate intermediate filament proteins; sequence characterization of two muscle proteins of a nematode. EMBO J. **11**, 3221–3227.