The EMBO Journal vol.5 no.8 pp.1865–1875, 1986

Amino acid sequence and gene organization of cytokeratin no. 19, an exceptional tail-less intermediate filament protein

Bernhard L.Bader, Thomas M.Magin, Mechthild Hatzfeld and Werner W.Franke

Division of Membrane Biology and Biochemistry, Institute of Cell and Tumor Biology, German Cancer Research Center, D-6900 Heidelberg, FRG

Communicated by Werner W.Franke

We have isolated a cDNA clone from a bovine bladder urothelium library which encodes the smallest intermediate filament (IF) protein known, i.e. the simple epithelial cytokeratin (equivalent to human cytokeratin 19) previously thought to have mol. wt 40 000. This clone was then used to isolate the corresponding gene from which we have determined the complete nucleotide sequence and deduced the amino acid sequence of the encoded protein. This cytokeratin of 399 amino acids (mol. wt 43 893) is identified as a typical acidic (type I) cytokeratin but differs from all other IF proteins in that it does not show the carboxyterminal, non-α-helical tail domain. Instead it contains a 13 amino acids extension of the α-helical rod. The gene encoding cytokeratin 19 is also exceptional. It contains only five introns which occur in positions correspondiong to intron positions in other IF protein genes. However, an intron which in all other IF proteins demarcates the region corresponding to the transition from the α-helical rod into the non-α-helical tail is missing in the cytokeratin 19 gene. Using *in vitro* reconstitution of purified cytokeratin 19 we show that it reacts like other type I cytokeratins in that it does not form, in the absence of a type II cytokeratin partner, typical IF. Instead it forms 40–90 nm rods of 10–11 nm diameter which appear to represent lateral associations of a number of cytokeratin molecules. Our results demonstrate that the non-α-helical tail domain is not an indispensable feature of IF proteins. The gene structure of this protein provides a remarkable case of a correlation of a change in protein conformation with an exon boundary.

Key words: keratins/intermediate filaments/multigene families/protein domains/epithelia

Introduction

Intermediate-sized filaments (IF) of diameter 7–12 nm are cytoskeletal structures which are formed by members of a large multigene family of proteins (e.g. Franke *et al.*, 1982b; Lazarides, 1982; Weber and Geisler, 1984; Steinert *et al.*, 1985b). Structural and biochemical studies as well as conformational probability analyses of amino acid sequences have led to models of organization of IF proteins in which the following major domains can be recognized: (i) a central rod domain comprising some 310 amino acids which is predominantly arranged in coiled-coiled α-helices, with at least two characteristic, albeit relatively short interruptions; (ii) an amino-terminal, non-α-helical head domain of variable length; and (iii) a carboxy-terminal tail region which is also not α-helical and shows extreme length differences between different IF proteins (e.g. Geisler and Weber, 1982; Geisler *et al.*, 1982; Hanukoglu and Fuchs, 1982,1983; Crewther *et al.*, 1983; Quax *et al.*, 1983; Steinert *et al.*, 1983,1985b).

From biochemical experiments, such as reconstitution of IF from purified polypeptides *in vitro*, immunological cross reactivities and amino acid comparisons, the IF proteins have been divided into three major subfamilies, the acidic (type I) and basic (type II) cytokeratins, which are specific for epithelial cells (Franke *et al.*, 1978; Sun and Green, 1978), and the six non-epithelial IF proteins, i.e. vimentin, desmin, glial filament protein and the three neurofilament polypeptides NF-L, NF-M and NF-H. While the various non-epithelial IF proteins exhibit relatively high degrees of homologies with each other (50–70%), homologies between the three major subfamilies are <30% (Hanukoglu and Fuchs, 1983; Weber and Geisler, 1984). A fourth subfamily related to IF proteins, as defined by partial sequence homologies, appears to be represented by the nuclear lamins A and C, i.e. karyoskeletal proteins forming the cortical lamina of the nucleus (McKeon *et al.*, 1986).

Sequence homologies common to the diverse IF proteins are generally restricted to certain portions of the α-helical rod, the most prominent being the canonical 'consensus sequence' at the end of the rod region (Geisler and Weber, 1982; Hanukoglu and Fuchs, 1982,1983; Quax *et al.*, 1983; Weber and Geisler, 1984; Steinert *et al.*, 1983; McKeon *et al.*, 1986). Analyses of the organization of the genes coding for the various IF proteins have shown that they contain a number of introns in corresponding positions, most of which are also related to the rod region (Quax *et al.*, 1983,1985; Lehnert *et al.*, 1984; Marchuk *et al.*, 1984; Balcarek and Cowan, 1985; Johnson *et al.*, 1985; Krieg *et al.*, 1985; Rieger *et al.*, 1985; Tyner *et al.*, 1985), suggesting that they are derived from a common ancestral gene (Franke *et al.*, 1978).

Among the IF proteins the cytokeratins are exceptional in that they appear as obligatory heteropolymers of tetrameric subunits containing two acidic (type I) and two basic (type II) polypeptides (Ahmadi *et al.*, 1980; Woods, 1983; Quinlan *et al.*, 1984, 1985; Hatzfeld and Franke, 1985; Parry *et al.*, 1985; Steinert *et al.*, 1985b). They are a very complex group of IF proteins (Franke *et al.*, 1981a), comprising in human tissues 19 different polypeptides, which are expressed in cell type specific patterns (Moll *et al.*, 1982; Sun *et al.*, 1984). Similar complexities and patterns of tissue-specific expression have been reported for other mammalian species (e.g. Franke *et al.*, 1982b; Schiller *et al.*, 1982; Tseng *et al.*, 1982).

Of the various cytokeratins, one acidic (type I) polypeptide is characterized by an especially high mobility in SDS–polyacrylamide gel electrophoresis (SDS–PAGE), corresponding in many gel systems to an apparent mol. wt of ~40 000. This polypeptide, which is encoded by a specific mRNA (Kim *et al.*, 1983; Magin *et al.*, 1983), is expressed in diverse kinds of epithelia of animals and humans such as intestinal mucosa (Franke *et al.*, 1981b, 1982b; Moll *et al.*, 1982), mesothelium (Wu *et al.*, 1982), mammary gland (Moll *et al.*, 1982; Bartek *et al.*, 1985), urothelium (Moll *et al.*, 1982; Rheinwald and O'Connell, 1985),

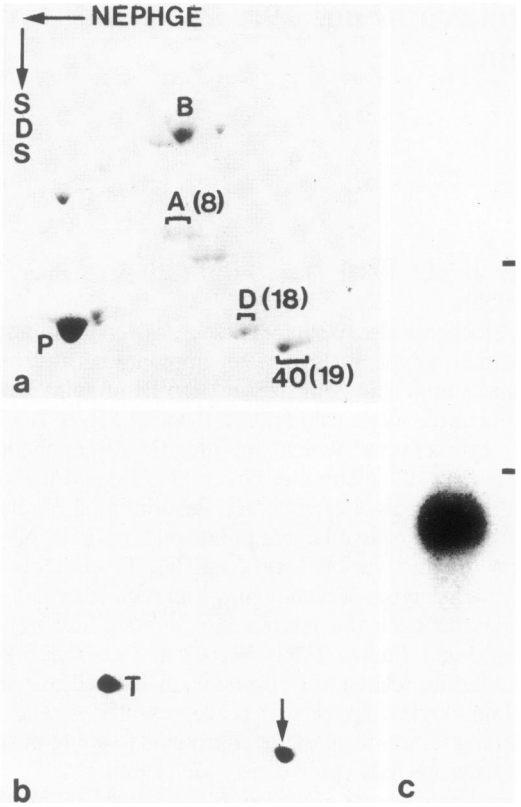

Fig. 1. Identification of a cDNA clone coding for bovine cytokeratin 19 (40 kd) by *in vitro* translation of mRNA hybrid-selected by binding to clone pKB19[1] **(a,b)** and by Northern blot analysis **(c)**. **(a)** Coomassie blue-stained gel after two-dimensional gel electrophoresis (NEPHGE, non-equilibrium pH gradient electrophoresis; SDS, SDS – PAGE used in second dimension gel electrophoresis) of cytoskeletal proteins from bovine bladder urothelium. Symbols A, D and '40' denote cytokeratins A (no. 8), D (no. 21 of the bovine catalog of Schiller *et al.*, 1982) corresponding to cytokeratin 18 of the human catalog of Moll *et al.* (1982) and the '40 kd' cytokeratin corresponding to no. 19 of human catalog. B and P, bovine serum albumin and yeast phosphoglycerokinase, used as markers in co-electrophoresis. **(b)** Fluorograph showing [^{35}S]methionine-labeled products of *in vitro* translation of mRNA selected by hybridization of urothelial poly(A)$^+$ RNA to clone pKB19[1]. The arrow denotes the position of cytokeratin 19 (40 kd), T is an endogenous product of the reticulocyte lysate. **(c)** Autoradiograph, showing Northern blot analysis of poly(A)$^+$ RNA from bovine bladder urothelium hybridized with clone pKB19[1]. The estimated size of the mRNA coding for cytokeratin 19 (40 kd) is 1.45 kb (positions of *Escherichia coli* rRNAs used as markers are indicated by bars).

certain embryonic epithelia (Sun *et al.*, 1985), some stratified epithelia and carcinomas (Moll *et al.*, 1982,1983; Schmidt *et al.*, 1982; Quinlan *et al.*, 1985) and in several epithelial cell culture lines (Wu and Rheinwald, 1981; Moll *et al.*, 1982; Morris *et al.*, 1985; Banks-Schlegel and Quintero, 1986). This cytokeratin is of particular interest in that its small size suggests that it represents a 'near the minimum size intermediate filament protein' (Wu and Rheinwald, 1981; Kaufmann *et al.*, 1985). To avoid problems of differences of mol. wt estimations from different SDS – PAGE systems as well as differences in the numbers of cytokeratin polypeptides identified in different species, we refer, in the following, to this polypeptide as 'cytokeratin 19 (40 kd)', according to the human cytokeratin catalog (Moll *et al.*, 1982; this polypeptide is identical to bovine and murine cytokeratin no. 22 of Schiller *et al.*, 1982). Here we present the sequence of the protein and its gene and show that it is an exceptional kind of simple epithelial type I cytokeratin.

Results

Identification of DNA clones

Poly(A)$^+$ RNA from urothelium of bovine bladder, a tissue known to be rich in cytokeratin no. 19 (40 kd) and mRNA encoding it (Moll *et al.*, 1982; Schiller *et al.*, 1982; Magin *et al.*, 1983; Rheinwald and O'Connell, 1985), was used to prepare a cDNA library in pUC8 plasmids. This library was screened, under conditions allowing cross-hybridization with other type I cytokeratin mRNAs, with a cDNA probe containing sequences encoding a bovine epidermal type I cytokeratin, i.e. component VII (pKBVII[1]; Jorcano *et al.*, 1984b,1984c; no. 16 of the bovine cytokeratin catalog of Schiller *et al.*, 1982) which is equivalent to human epidermal cytokeratin no. 14 (Hanukoglu and Fuchs, 1982; Jorcano *et al.*, 1984b,1984c; Marchuk *et al.*, 1984). Clones identified by this procedure were characterized by *in vitro* translation of hybrid-selected urothelial mRNA and found to encode specifically a cytokeratin polypeptide that co-migrated, on two-dimensional gel electrophoresis, with authentic bovine cytokeratin 19 (Figure 1a and b). This cDNA clone was then used to select bovine genomic clones which were identified in the same way (not shown). Clone pKB19[1] was also used for Northern blot analyses on total urothelial poly(A)$^+$ RNA and revealed only one band corresponding to a size of ~ 1.45 kb (Figure 1c). This value is slightly lower than the 1.585 kb reported for mRNA encoding human cytokeratin 19 by Eckert and Green (1984).

Nucleotide sequences and animo acid sequences deduced therefrom

Figure 2 presents the nucleotide sequence of the coding part of the genomic clone λKB19[6], together with the 5′ non-coding portion and 969 nucleotides upstream of the putative cap site, the 3′ non-coding portion and 222 nucleotides downstream of the polyadenylation signal. This Figure also includes the complete sequences of four of the five introns present; intron 1 which is very long, >1.7 kb, is only shown in part. The positions of the introns were identified by comparison with the cDNA clone and by S1 nuclease mapping. The latter procedure was also used to identify the start of the mRNA. The nucleotide sequence of the cDNA clone was fully identical with that of the corresponding portion of the genomic clone λKB19[6].

The 3′ non-coding region of the mRNA is relatively short and contains only 131 nucleotides, including the stop codon and the polyadenylation site which was identified at the beginning of the poly(A) stretch of 86 residues present in the cDNA clone pKB19[1] (denoted in Figure 2 by the upward arrowhead following the polyadenylation site; for details see Bader, 1986). The total length of the mRNA, as estimated from the nucleotide sequences, is 1473 nucleotides including the 86 residues of the poly(A) stretch recoverd in the cDNA. This is in good agreement with the value of 1.45 kb determined by Northern blot analysis (Figure 1c).

The coding region of the mRNA comprises 1197 nucleotides, i.e. significantly more than the estimate of 1090 nucleotides by Eckert and Green (1984) for human cytokeratin 19 mRNA which, however, was based on the assumption of a total protein mol. wt of 40 000 (cf. Franke *et al.*, 1981b; Fuchs and Green, 1981; Wu and Rheinwald, 1981; Schiller *et al.*, 1982; Tseng *et al.*, 1982; Bartek *et al.*, 1985). In fact, however, the polypeptide encoded by this gene contains 399 amino acids of a total mol. wt of 43 893, including the first methionine which is probably not preserved in the mature polypeptide. This value, which is close to mol. wt estimations from certain gradient gel systems (see Figure 7) and the system used by Banks-Schlegel and Quintero (1986), suggests that the protein tends to run anomalously fast

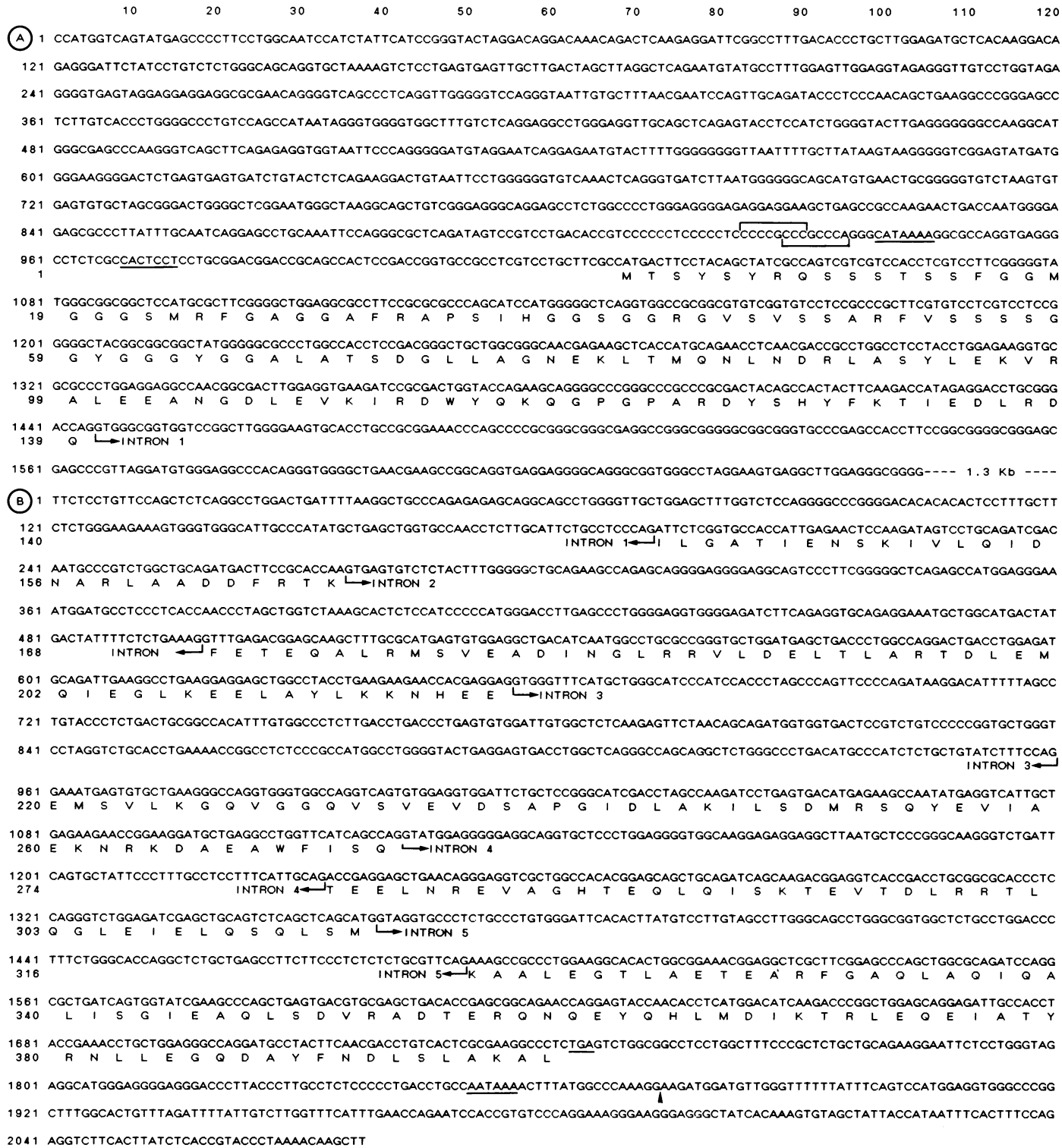

Fig. 2. Nucleotide sequence of bovine cytokeratin 19 gene and amino acid sequence deduced therefrom. The complete sequence of the gene encoding cytokeratin 19 (40 kd) is shown, except for the very long intron 1 which is only partially included. Adjacent upstream and downstream regions are also shown. The amino acids encoded by the exons are indicated in single-letter code. Division of the cytokeratin 19 gene into two blocks (**A** and **B**), divided within the long intron 1, is to facilitate reading. Nucleotides are numbered from 1 to 1664 in **A** and from 1 to 2078 in **B**. The introns are numbered and the intron boundaries are indicated by arrows pointing towards the center of the intron. The putative promoter sequence CATAAAA, the cap site sequence, the stop codon and the polyadenylation signal are underlined. The brackets in front of the CATAAAA sequence denote two interdigitated GC boxes. The arrowhead in the third line from the bottom denotes the end of the corresponding cDNA clone pKB19[1] as identified by the poly(A) stretch of 86 residues contained in the cDNA clone.

in certain SDS−PAGE systems, probably due to the special amino acid composition of the head portion (see below) and/or retention of some secondary structure.

A comparison of the deduced amino acid sequence of cytokeratin 19 with those of other IF proteins (Figures 3 and 4) revealed that it is a typical member of the acidic (type I) cytokeratin

Fig. 3. Comparison of the amino acid sequence of cytokeratin 19 (40 kd) with those of type I and II cytokeratins and vimentin. Amino acid sequences (one letter code) of the IF proteins, as deduced from the corresponding nucleotide sequences, have been aligned for maximal homology (insertions introduced for this purpose are denoted by horizontal bars): human cytokeratin 14 (H14, 50 kd) as a representative of the acidic (type I) cytokeratin subfamily (data from Marchuk *et al.*, 1984,1985), partial sequence of bovine cytokeratin 8 (B8, A) as a representative simple epithelial basic (type II) cytokeratin (data from Magin *et al.*, 1986), human cytokeratin 6 (H6, 56 kd) as a representative of epidermal basic (type II) cytokeratins (data from Tyner *et al.*, 1985), and hamster vimentin (HA VIM) as a representative of the non-epithelial IF proteins (data from Quax *et al.*, 1983,1985). Bold-faced letters denote amino acids identical in bovine cytokeratin 19 (40 kd) and at least one of the other IF proteins. The downward arrow demarcates the start and upward arrow the end of the α-helical rod. The lines and dots above the sequence blocks indicate the extent of the coiled-coil subdomains C1a, C1b, C2 of the rod domain and the α-helical extension (HE) of 13 amino acids. The dots represent positions *a* and *d* of the heptade convention to maximize coiled-coil configuration. The rod domain reveals two non-α-helical interruptions of 14 and 20 amino acids, respectively. The two arcs at positions 266 and 326 of cytokeratin 19 indicate probable sites of skip and/or reversal of polarity in the coiled-coil. Amino acid motifs in the head portion of cytokeratin 19 that are also recognized in the head regions of other IF proteins are underlined. Note absence of homologies in the carboxy-terminal tail region. The 23 residues of the amino-terminal amino acid sequences of the mol. wt 39 000 chymotryptic fragment from cytokeratin 19 (cf. Figure 7), as present in IF reconstituted from purified bovine cytokeratins 8 and 19, is indicated by a bracket, designated 'P', with an arrowhead pointing to the amino terminus of this fragment. Asterisks denote carboxy termini.

subfamily, confirming previous classifications based on nucleic acid cross-hybridization at reduced stringency, peptide mapping, antibody epitope patterns and *in vitro* reconstitution experiments (e.g. Fuchs *et al.*, 1981; Schiller *et al.*, 1982; Tseng *et al.*, 1982; Kim *et al.*, 1983; Hatzfeld and Franke, 1985). Probability analyses of the conformation of this protein by the methods of Chou and Fasman (1978; data not shown) and Garnier *et al.* (1978) showed many regions with α-helical character (Figure 5a) and displayed only three non-α-helical regions: the head portion of 71 amino acid residues and two relatively short interruptions demarcating coils 1a, 1b and 2 of the rod portion which comprises 315 amino acids. Conventional heptade patterns indicative of coiled-coil configuration (McLachlan, 1978) are also included in Figure 3.

Cytokeratin 19 is different from all other IF proteins in that it has no typical tail domain (Figure 5b presents the comparison with a cytokeratin of the same, i.e. type I, subfamily of the same species). Instead of the abrupt halt in α-helical character and the heptad frame which is found in all other IF proteins after the TYR(X)LLEG$^{E}_{Q}$ consensus sequence and defines the transition from the rod to the tail domain, cytokeratin 19 shows a continuation of the α-helical character and the heptad frame throughout the short carboxyterminal stretch of 13 amino acids (Figure 5c,d). We conclude that cytokeratin 19 does not possess a non-α-helical tail domain but instead ends with a short α-helical extension of the rod (Figure 5a – d). Interestingly, this α-helical extension contains a terminal oligopeptide motif, DLSLAKAL, which is very similar to certain evolutionary stable motifs in the coiled-coil α-helical rod portions of the myosin heavy chain (for example, SLAKA and ELTLAKA; cf. McLachlan and Karn, 1982) and some non-epithelial IF proteins such as glial filament protein (DFSLAGAL: Weber and Geisler, 1984; Balcarek and Cowan, 1985).

Amino acid sequence homologies of cytokeratin 19 (40 kd) with cytokeratins of the basic (type II) subfamily or with non-epithelial IF proteins are rather low and essentially restricted to the rod. For example, the rod of bovine cytokeratin 19 has 87 amino acids in common (28%) with human cytokeratin 6 (mol. wt 56 000; Hanukoglu and Fuchs, 1983; Tyner *et al.*, 1985) and 104 residues (33%) with hamster vimentin (Quax *et al.*, 1983). Typically, these homologies between different IF protein subfamilies tend to concentrate in the center of coil 1a as well as near the beginning and end of coil 2 (Figure 3). In addition, several scattered homologies are recognized which contain a high percentage of charged amino acids (53 homologous positions of the rod are occupied by charged amino acids). In contrast, homologies between bovine cytokeratin 19 and other members of the same (type I) subfamily are usually very extensive but are also concentrated in the rod domain. Figure 4a presents several examples of high homology (70 – 73% identity) between rod sequences of the simple epithelial cytokeratin 19 of bovine origin with epidermal cytokeratins of human (no. 14; Marchuk *et al.*, 1984), bovine (component VIb; Rieger *et al.*, 1985) and amphibian (XK81 of mol. wt 47 000; Jonas *et al.*, 1985; see also Hoffmann and Franz, 1984) origin. Somewhat surprisingly, the number of identical amino acids was much lower when cytokeratin 19 was compared with another simple cytokeratin of the same (type I) subfamily, i.e. murine (Singer *et al.*, 1986) and human cytokeratin 18 of mol. wt 45 000 (~45% identity in the rod portion; comparison with a published partial sequence of human cytokeratin 18 is shown in Figure 4b).

The head portion of cytokeratin 19 is relatively short (71 amino

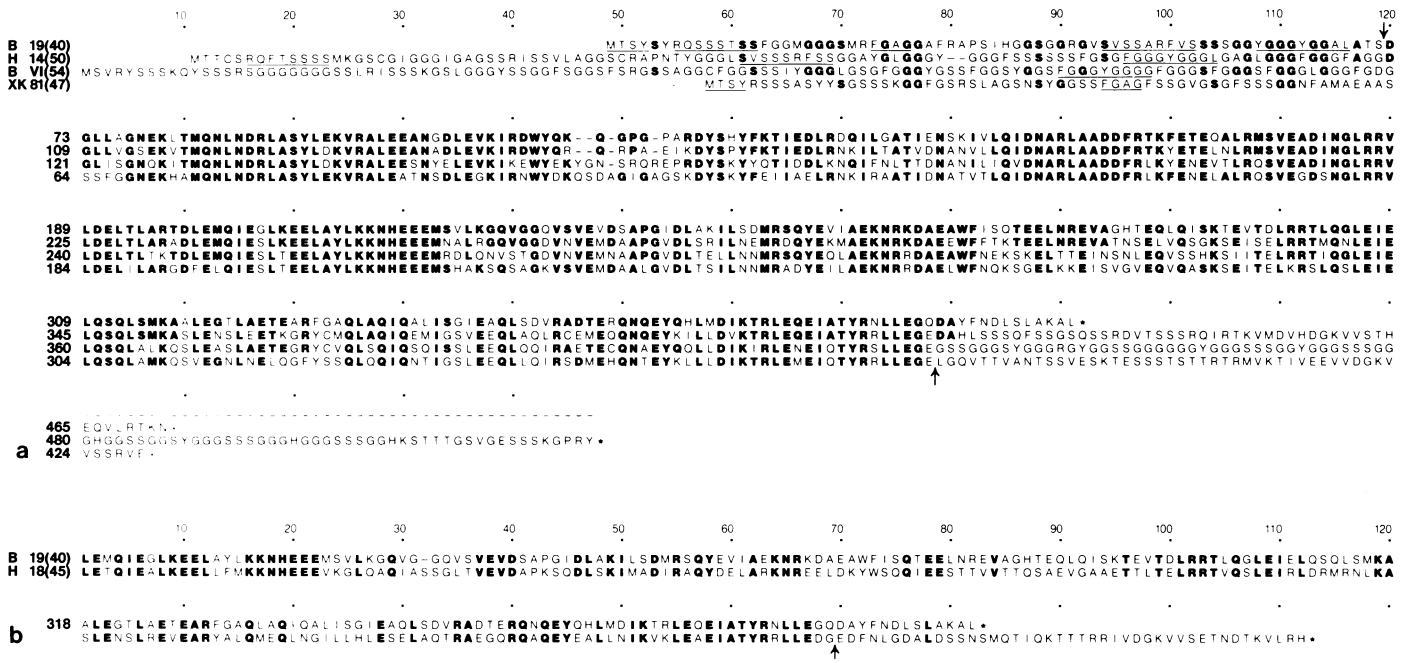

Fig. 4. Comparison of the amino acid sequences of cytokeratin 19 (40 kd) with several representatives of the acidic (type I) cytokeratin subfamily. Symbols are as in Figure 3. **(a)** Comparison with type I cytokeratins of epidermal type from diverse species: human cytokeratin 14 (H 14, 50 kd; data from Marchuk *et al.*, 1985), bovine epidermal cytokeratin VI (54 kd, data from Rieger *et al.*, 1984,1985), epidermal type cytokeratin XK 81 (47 kd), expressed in certain embryonal stages of *Xenopus laevis* (data from Jonas *et al.*, 1985). **(b)** Comparison with partial sequence of the simple epithelial cytokeratin 18 from human liver (H18, 45 kd; Romano *et al.*, 1986). Alignment and symbols as in Figure 3.

acid residues) but there are other IF proteins with even shorter heads (Figure 4a) such as the glial filament protein (Weber and Geisler, 1984; Balcarek and Cowan, 1985) and the epidermal mol. wt 47 000 cytokeratin of *Xenopus* termed XK81 (Jonas *et al.*, 1985). While the head does not display sequence homologies common to all IF proteins it shows the general predominance of basic amino acids, mostly arginines (in fact the head of cytokeratin 19 does not contain a single negatively charged amino acid) as well as several sequence features that are also found in some epidermal type I cytokeratins. These include an identical amino terminus MTSY and a FGAG motif in *Xenopus* cytokeratin XK81, similar hydroxyamino acid-rich motifs such as $RQS(S,T)_5$ and $SVSS^A_S RF^V_S S$ in human cytokeratin 14, and the notorious glycine-rich oligopeptides such as $(Y,F,L)GGG^Y_F GG^A_G$ (Figures 3 and 4a). Similar oligopeptide repeats rich in glycine and aromatic amino acids occur in heads and/or tails of several basic (type II) cytokeratins (Hanukoglu and Fuchs, 1983; Jorcano *et al.*, 1984a; Steinert *et al.*, 1985a,1985b; Tyner *et al.*, 1985). The functional importance of such sequence motifs in the head is not clear.

Intron positions in the cytokeratin 19 gene

Figure 2 presents the introns of the gene encoding bovine cytokeratin 19, and the positions of these introns are compared with those of introns determined in other IF genes in Figure 6. While all other IF protein genes sequenced so far have seven or eight introns (Quax *et al.*, 1983,1985; Lehnert *et al.*, 1984; Marchuk *et al.*, 1984,1985; Balcarek and Cowan, 1985; Johnson *et al.*, 1985; Krieg *et al.*, 1985; Rieger *et al.*, 1985), the cytokeratin 19 gene contains only five introns which vary greatly in size (from 111 to 1720 nucleotides). The sequences surrounding the exon−intron boundaries fit the 'consensus sequence' requirements as defined by Mount (1982). The exons demarcated by introns fall into a relatively narrow range of sizes (exons 2−5 comprise 83, 157, 162 and 126 nucleotides, respectively) as is found in many eukaryotic genes (Naora and Deacon, 1982), other type I cytokeratins included (Rieger *et al.*, 1985). All of the introns present in the bovine cytokeratin 19 gene correspond in position to introns described in other IF protein genes (Figure 6a−e). However, unlike the other IF protein genes, the gene for cytokeratin 19 does not possess the common intron located right after the consensus amino acid sequence, i.e. a position denoting the change from α-helical rod to non-α-helical tail (Figure 6f). Of course, introns located in positions corresponding to tail regions of the protein are absent in the more distant tail portions (Figure 6g−j).

Candidates for upstream regulatory elements

We searched the ~1 kb DNA sequence upstream of the cap site for possible regulatory elements. Immediately preceeding a CATAAAA sequence, which obviously is the functional equivalent to the canonical TATA box (for review see Breathnach and Chambon, 1981), we noted two interdigitated GGGCGGG elements ('GC boxes') as they are known from several viral genomes and some eukaryotic genes (Dynan and Tijan, 1985), which are embedded in a 30 nucleotide long stretch of purine−pyrimidine asymmetry. In position −375 to −333 (positions 595−637 of Figure 2A) a sequence was found which contained two blocks of high homology to bovine papilloma enhancer sequences (block 1 with 9 out of 10 nucleotides identical, block 2 with 15 out of 19 identical; cf. Lusky *et al.*, 1983).

Determination of protein structure protected against chymotryptic digestion

To determine the region of cytokeratin 19 that is protected against proteolytic attack by its inclusion in the α-helical coiled-coil configuration we purified bovine cytokeratins 8 and 19 by a combination of DEAE-cellulose anion exchange chromatography and

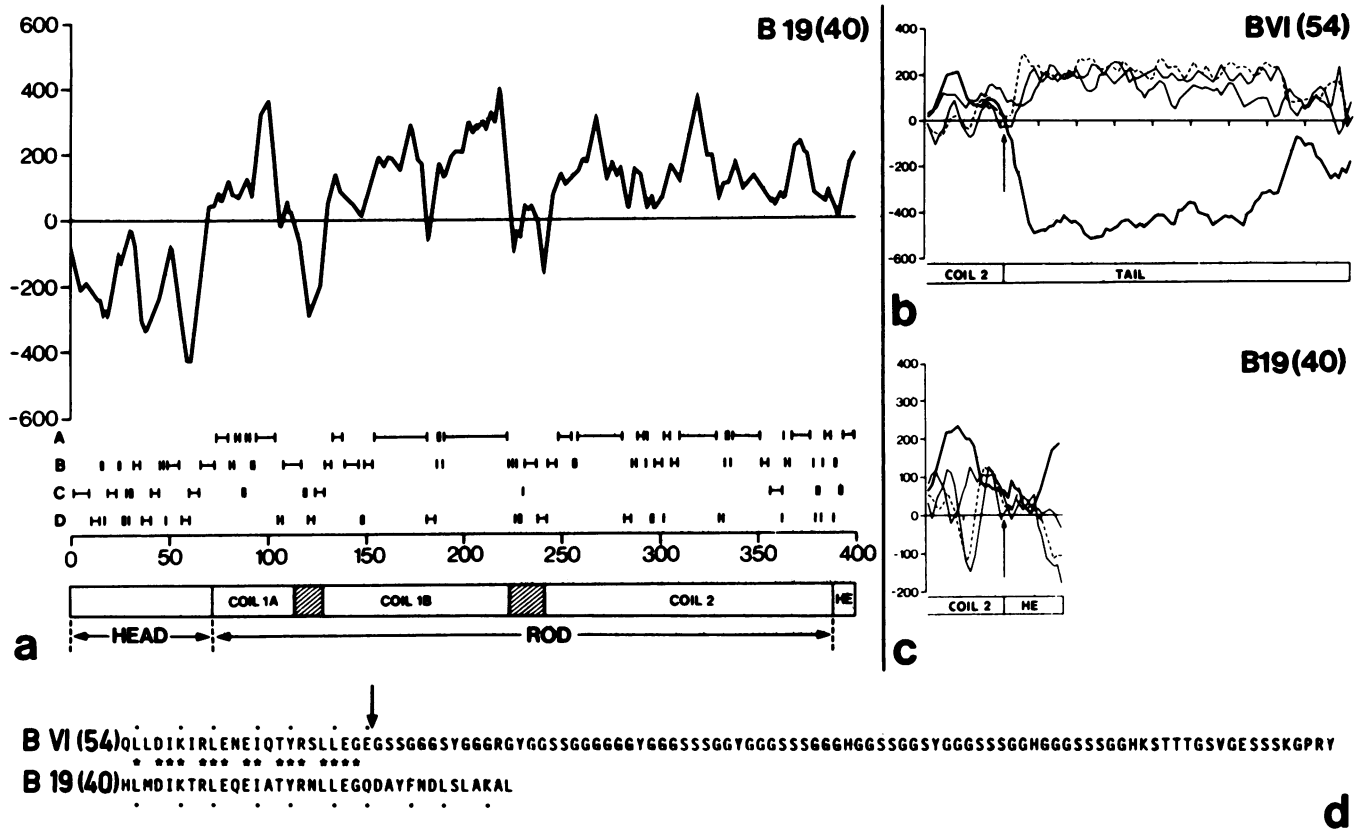

Fig. 5. Secondary structure prediction of bovine cytokeratin 19 in comparison with other intermediate filament proteins. **(a)** The upper part shows an α-helix probability profile for the amino acid sequence of cytokeratin 19. The program of Garnier *et al.* (1978) was used with a sliding average of eight amino acids (plotted as function of amino acid position). The values on the ordinate present relative units of probability of α-helical conformation in information theoretical units (centinats). The lower part shows the relative probabilities for α-helix (A), β-sheet (B), β-turn (C) and random coil (D) conformation as linear arrays. Positions of amino acid residues in cytokeratin 19 are given by the numbers at the bottom line. The structural domains of cytokeratin 19 are presented, in the same scale, in the bottom panel. Head, amino-terminal domain, coil 1A, 1B, 2 represent the α-helical subdomains of the rod, interrupted by two considerable non-α-helical regions (hatched boxes). 'HE' denotes the α-helical carboxy-terminal extension of 13 amino acids. **(b,c)** Secondary structure prediction of the ends of coil 2 and the adjacent carboxy-terminal regions of bovine cytokeratins 19 and VI (using the sequence data from Rieger *et al.*, 1985). The thick black line represents the α-helical probability, whereas the other three lines represent probabilities for the non-α-helical secondary structures as B, C, D in **a**. The arrow denotes the end of the rod, where in bovine cytokeratin VI as in other IF proteins the α-helicity of coil 2 drastically decreases and changes into the non-α-helical carboxyterminal tail **(b)**. In contrast, in cytokeratin 19 α-helicity does not end in this position but is extended by another 13 amino acid residues (α-helical extension, HE; **c**). **(d)** Amino acid sequences corresponding to the regions of cytokeratin B VI (54) and B 19 (40) shown in **(c)**. The arrow demarcates the end of coil 2 (as in **b,c**). Dots above and below amino acid sequences indicate positions *a* and *d* of the heptade convention for coiled-coil arrangement. Asterisks denote identical amino acids.

reverse phase h.p.l.c. (Figure 7) and reconstituted IF *in vitro* (Hatzfeld and Franke, 1985). The IF reconstituted from purified cytokeratins 8 and 19 were then digested with chymotrypsin (see Materials and methods), as has been done with other kinds of reconstituted IF or protofilaments (e.g. Steinert *et al.*, 1980; Geisler *et al.*, 1982; Kaufmann *et al.*, 1985). The major protected fragment showed, on SDS – PAGE, an apparent mol. wt value of ~39 000 (Figure 7b, lane 2), using a gel system in which intact cytokeratin 19 appeared with an apparent mol. wt of ~44 000 (Figure 7b, lane 1); see also Banks-Schlegel and Quintero, 1986). Determination of the amino acid sequence of this fragment (see Materials and methods) identified its amino terminus as the valine at a position 19 residues before the start of the rod domain, as indicated in Figure 3. The total of mol. wt of this fragment as calculated from the amino acid sequence was 38 768 which indicated that most of the head portion has been removed and the tyrosin and phenylalanine residues of the carboxyterminal α-helical extension of the rod had been protected.

Structures of cytokeratin 19 formed in vitro

It is generally accepted that cytokeratin IF are obligatory heteropolymers and require the formation of heterotypic subunit complexes containing equal numbers of type I and type II cytokeratin polypeptides (for references see Introduction; for cytokeratin 19 see Hatzfeld and Franke, 1985). To examine whether the exceptional feature of cytokeratin 19, i.e. the absence of a non-α-helical tail, is correlated with a different assembly behaviour we have studied by electron microscopy the *in vitro* reconstitution and assembly of purified bovine cytokeratin 19 under various conditions, including a considerable range of ionic strengths of the re-assembly buffers used (see Materials and methods). While in low ionic strength buffers such as 2 mM Tris-HCl (pH 8.0) and 10 mM Tris-HCl (pH 7.5) only 40 – 50 nm threads of ~2 nm diameter were seen, sometimes revealing a polar globule (data not shown), as previously described for other cytokeratins (Hatzfeld and Franke, 1985), higher ionic strength resulted in a pronounced tendency to form thicker rods as visualized both by metal shadow-cast and negatively stained prepara-

Fig. 6. Positions of introns in genes encoding different intermediate filament proteins. **(a – j)** The nucleotide and amino acid sequences surrounding the introns (insertion site denoted by triangle) are shown. The specific numbers of the introns are given in the triangles. Alignments and symbols are as in Figures 3 and 4. Genes compared are: bovine cytokeratin 19 (40 kd); murine epidermal type I cytokeratin of mol. wt 59 000 (M59; data from Krieg *et al.*, 1985); bovine epidermal type I cytokeratin VI (BVI; 54 kd; data from Rieger *et al.*, 1985); human epidermal type I cytokeratin 14 (H14; 50 kd; data from Marchuk *et al.*, 1984,1985); hamster vimentin (HA VIM; data from Quax *et al.*, 1983); human epidermal type II cytokeratin 1 (H1; 67 kd; data from Johnson *et al.*, 1985); human type II cytokeratin 6 (H6; 56 kd; Tyner *et al.*, 1985); bovine epidermal type II cytokeratin Ib (BIb; 68 kd; partial sequence from Rieger *et al.*, 1985). Positions of introns in hamster desmin and murine glial filament protein are essentially as in vimentin (Balcarek and Cowan, 1985; Quax *et al.*, 1985). Note corresponding intron positions of introns 1 – 5 of cytokeratin 19 gene and other intermediate filament protein genes **(a – e)** whereas the cytokeratin 19 gene has no intron in the rod-tail transition position which is denoted by an intron in all other IF protein genes **(f)**. Introns corresponding to the tail regions vary greatly between different IF protein genes **(g – j)**.

tions (Figure 8). Measurements of structures obtained in 50 and 100 mM Tris-HCl buffer solutions showed an average diameter (determined from negative stainings) of ~10.5 nm and lengths varying from 40 to 90 nm, revealing a biphasic length distribution with mean peak values of 45 and 72 nm, with both populations occurring in similar frequencies. In control experiments in which stoichiometric amounts of bovine cytokeratin 8 were added normal-looking IF structures were found as previously described (Hatzfeld and Franke, 1985). This shows that cytokeratin 19 alone does not form IF but assembles into short rod-like aggregates of about the width of an IF, suggesting that many molecules associate laterally but do not polymerize linearly.

Discussion

Knowledge of sequences and gene arrangements of cytokeratins was restricted so far to epidermal type keratins (for partial characterization of the gene encoding the simple epithelial cytokeratin A of mouse, equivalent to human cytokeratin 8, see Vasseur *et al.*, 1985). The sequence analysis of clone λKB19[6] now presents the organization of a simple epithelial cytokeratin gene. The cytokeratin 19 encoded by this gene is the smallest known IF protein and is characteristic of diverse vertebrate epithelia of species as taxonomically distant as mammals and amphibia (Franz *et al.*, 1983). From the amino acid sequence of the normal-sized rod (315 amino acids) this small cytokeratin is identified as a typical member of the acidic (type I) subfamily of cytokeratins and is highly homologous to epidermal type I keratins. While the heads and the tails of most epidermal cytokeratins are characterized by oligopeptide repeats of motifs rich in glycine combined with aromatic amino acids and/or serine (for a different kind of epidermal type I keratin tails see Hanukoglu and Fuchs, 1982; Jorcano *et al.*, 1984b) only a single, rudimentary GYGGGYGG sequence is seen in the head portion of cytokeratin 19. In other aspects, the relatively short head portion of cytokeratin 19 is similar to the heads of other IF proteins, in that it is very positively charged, rich in hydroxyamino acids and shows a number of oligopeptide motifs that are also found in epidermal cytokeratins (Figures 3 and 4).

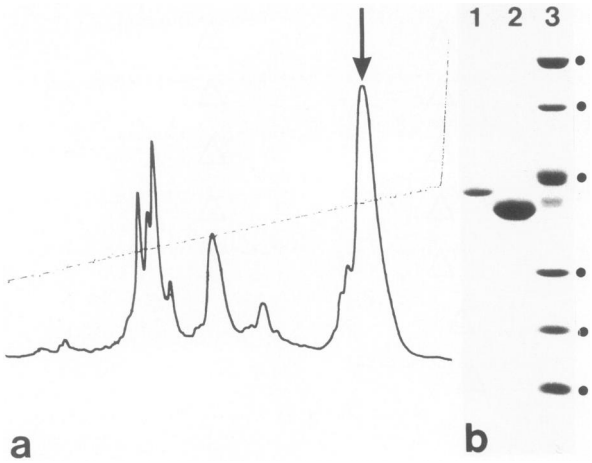

Fig. 7. Purification of the major bovine cytokeratin 19 fragment obtained by limited proteolysis of intermediate filaments with chymotrypsin. **(a)** H.p.l.c. elution profile of chymotryptic fragments from cytokeratin polypeptides nos. 8 and 19 as obtained after limited digestion of intermediate filaments reconstituted from purified polypeptides *in vitro*. The abscissa indicates the elution time, elution was monitored at 206 nm (ordinate). The arrow denotes the 39 kd fragment from cytokeratin 19. **(b)** SDS – PAGE of purified cytokeratin polypeptide 19 **(lane 1)** and the 39-kd fragment **(lane 2**; same material as denoted by the arrow in **a)**. Molecular weight markers are **(lane 3**, dots from top to bottom): phosphorylase B (97 400), BSA (66 000), ovalbumin (45 000), carbonic anhydrase (29 000), soybean trypsin inhibitor (20 000) and lysozyme (14 300).

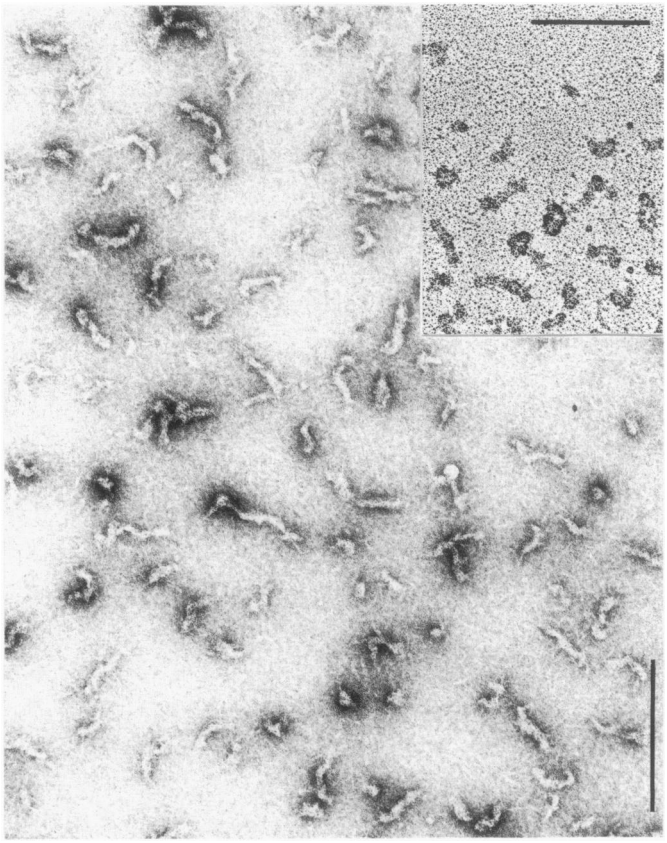

Fig. 8. Electron micrographs (negative staining; the insert in the upper right presents a metal shadow-cast preparation of the same material) of structures formed by h.p.l.c.-purified bovine cytokeratin 19 reconstituted and re-assembled *in vitro* (final dialysis buffer in the experiment shown here: 50 mM Tris-HCl, pH 7.5, 1 mM dithiothreitol). Most of the material is contained in rods 40 – 90 nm long of 10 – 11 nm diameter which, in histograms, display a bimodal distribution. These structures are interpreted as lateral aggregates of a number of polypeptides. Occasional 2 – 3 nm rods of single polypeptides or oligomers are seen in the background (e.g. in the upper part of the insert). Bars, 200 nm.

The principal feature that distinguishes cytokeratin 19 from all other IF proteins, including other type I cytokeratins, is the absence of a non-α-helical tail portion. Therefore, cytokeratin 19 presents the exceptional example of an IF protein which by its mere existence demonstrates that the presence of a non-α-helical tail is not an indispensable feature of IF proteins. This is in line with the recent conclusions of Kaufmann *et al.* (1985) who removed enzymatically the last 27 amino acids from the carboxy terminus of chicken gizzard desmin and found that this 'half tail-clipped' molecule was still able to polymerize into IF *in vitro*. The absence of a non-α-helical tail in cytokeratin 19 also does not result in considerably different assembly properties. Like other type I cytokeratins, cytokeratin 19 requires the presence of stoichiometric amounts of a complementary type II cytokeratin to form IF, indicating that the requirement of heterotypic subunit complexes is not dependent on sequences or features located in the tail region. The existence of the tail-less cytokeratin 19 also shows that neither the glycine-rich oligopeptides repeat nor the various evolutionarily conserved 'signal' motifs (e.g. VDGKVVS; cf. Jorcano *et al.*, 1984b; Jonas *et al.*, 1985; Quinlan *et al.*, 1985) are essential for the fundamental interaction with the corresponding basic (type II) cytokeratin complex partner (Franke *et al.*, 1983; Sun *et al.*, 1984,1985; Fuchs *et al.*, 1985) and subsequent IF formation. It is widely believed that the tail regions of IF proteins project laterally from the protofilament backbone formed by linear arrays of the rods and are involved in the cell type-specific functions of the individual kind of IF (for reviews see Steinert *et al.*, 1985a,1985b; Weber and Geisler, 1985). However, the short α-helical extension of cytokeratin 19 protrudes only very little, if at all, from the protofilament surface and is unlikely to contribute to cell type-specific functions in those epithelia that synthesize this protein.

As the diverse IF proteins share common principles of domain organization their genes display similar exon-intron patterns (Lehnert *et al.*, 1984; Marchuk *et al.*, 1985; Balcarek and Cowan, 1985; Johnson *et al.*, 1985; Krieg *et al.*, 1985; Quax *et al.*, 1985; Rieger *et al.*, 1985; Tyner *et al.*, 1985). All IF protein genes analyzed so far have been shown to contain seven or eight introns. The gene coding for cytokeratin 19 contains only five introns. These occur in positions identical to intron positions in several other IF protein genes, either in the same codon or shifted to a neighbouring codon. This conservation of intron positions throughout IF protein evolution and diversification indicates an origin from a common genomic element, at least for the rod portion. Similar observations of evolutionary retention of the positions of some but not all introns have been reported for other multigene families (for review see Gilbert, 1985), including cytoskeletal protein families such as the myosin heavy chains (Strehler *et al.*, 1985), tubulins (Gwo-Shu Lee *et al.*, 1983) and actins (Cooper and Crain, 1982; Fornwald *et al.*, 1982; Ueyama *et al.*, 1984).

While some of the exon – intron boundaries in IF protein genes can be related to regions of changes of conformational character (Lehnert *et al.*, 1984; Krieg *et al.*, 1985; Rieger *et al.*, 1985), a relationship to possible structural or functional domains is not obvious for other introns (Quax *et al.*, 1983,1985; Marchuk *et al.*, 1984; Johnson *et al.*, 1985; Rieger *et al.*, 1985). However, for one of the introns present in IF proteins the relationship to a protein domain boundary is without question, i.e. the intron

located at the transition from the α-helical coiled-coil rod into the non-α-helical tail domain, that is, two or three codons after the consensus sequence that defines the carboxy-terminal end of the rod. Now the gene encoding cytokeratin 19 provides the exceptional example of the negative formulation of this correlation, i.e. the absence of this characteristic intron in a protein that does not show a change of conformation in this region. This case therefore strongly supports the concept that exons code for protein domains and that many of the introns in today's genes define splice points of distinct genomic elements carrying conformational blocks or functional elements (Gilbert, 1978,1985; Doolittle, 1978; Blake, 1983; Craik *et al.*, 1983; Lonberg and Gilbert, 1985; Südhof *et al.*, 1985). The absence of an intron in this position of the cytokeratin 19 gene might be explained by intron loss during vertebrate evolution, similar to the situation of the rat proinsulin gene (Perler *et al.*, 1980). However, it is difficult to see how the loss of this landmark intron in a precursor IF protein gene would result in the formation of the short α-helical rod extension with an amino acid sequence unrelated to the positionally corresponding tail regions of any of the other IF proteins. Therefore, we think that the special IF protein gene structure of cytokeratin 19 might represent an ancestral arrangement from which other genes of this protein family have evolved by intron insertion, resulting in recombinations with the various kinds of tail structures, or might result from two independent processes, namely intron-mediated replacement of a tail-encoding exon by an exon encoding the α-helical extension (type A3 intron *sensu* Stone *et al.*, 1985), followed by intron loss.

The small cytokeratin 19 is of particular interest in studies of the regulation of cytokeratin expression and epithelial cell differentiation as it shows a remarkably rapid inducibility in various epithelial cells where it appears in addition to the pre-existing set of cytokeratin polypeptides. Examples of inductions include its increased synthesis in cultured human epidermal keratinocytes, conjuctival cells and various squamous cell carcinoma lines upon addition of vitamin A, retinyl acetate and/or retinoic acids and analog compounds (Fuchs and Green, 1981; Eckert and Green, 1984; Kim *et al.*, 1984), the increase in keratinocytes after transformation with SV40 (Morris *et al.*, 1985) and the apparently environmentally influenced appearance of this protein in certain tumors but not in others (Moll *et al.*, 1983; Fuchs *et al.*, 1985; Banks-Schlegel and Quintero, 1986). Therefore it is likely that the regulation of expression of cytokeratin 19 involves special regulatory elements which are not shared by other cytokeratin genes. In this context it is interesting to note that the cytokeratin 19 gene displays, in the putative promoter region, two interdigitated 'GC boxes', i.e. elements that bind the specific cellular transcription factor, Sp1 protein (Dynan and Tjian, 1985). This kind of regulatory element is known to occur in several viral promotor regions as well as in certain cellular genes but is not found in many other genes. Among the diverse IF protein genes sequenced so far the occurrence of 'GC boxes' has been detected in the promoter regions of human cytokeratin 14 (Marchuk *et al.*, 1985) but not in any of the other cytokeratin genes (Johnson *et al.*, 1985; Krieg *et al.*, 1985; Rieger *et al.*, 1985; Tyner *et al.*, 1985). A pronounced 'GC box' also exists in the desmin gene (Quax *et al.*, 1985) but not in the genes coding for vimentin and glial filament protein, despite the great similarity of these proteins (Quax *et al.*, 1983; Balcarek and Cowan, 1985). Clearly, studies of Sp1 binding and GC-box-governed expression are needed to examine whether these elements are really involved as 'regulatory sequences' in the cell type-specific expression of cytokeratin 19.

Materials and methods

cDNA

Procedures for preparation of RNA from bovine bladder urothelium for cDNA cloning and for the identification of recombinant cDNA clones have been described (Magin *et al.*, 1983,1986). To screen for cDNA clones encoding bovine cytokeratin 19, a gel-purified insert of a cDNA clone encoding a bovine epidermal acidic (type I) cytokeratin, i.e. component VII (pKB VII[1]; Jorcano *et al.*, 1984b,1984c), was used as hybridization probe. Therefore, conditions of lowered stringency, allowing cross-hybridization, were applied. Hybridization was performed overnight at 37°C in 50% formamide 5 × SSPE (1 × SSPE is 0.18 M NaCl, 10 mM NaH_2PO_4, pH 7.4), 5 × Denhardt's solution, 0.2% SDS, 100 μg/ml yeast tRNA. Filters were washed four times at room temperature with 2 × SSC, 0.2% SDS and twice at 45°C for 45 min each with 0.2% SDS, 0.1 × SSC (1 × SSC is 0.15 M NaCl, 0.015 M Na-citrate). Out of five positive clones, one (pKB19[1]) of ~0.6 kb was chosen for further experiments. The polypeptide encoded by this clone was identified by *in vitro* translation of hybrid-selected mRNA from bovine urothelium, using the rabbit reticulocyte system for translation and two-dimensional gel electrophoresis for protein separation as described (Magin *et al.*, 1983; Jorcano *et al.*, 1984b).

Isolation and characterization of genomic clones

We screened a bovine genomic library in λ-phages (Ruppert *et al.*, 1984) as previously described (Lehnert *et al.*, 1984) using nick-translated pKB19[1] as hybridization probe. Identification of genomic clones coding for cytokeratin 19 was as described above for clone pKB19[1]. The phage clone selected for further characterization (λKB19[6]) contained an insert of ~15 kb.

Recombinant phages and plasmids containing subclones thereof were digested with several restriction endonucleases. Fragments were separated by electrophoresis on agarose gels and blotted on nitrocellulose filters using the bidirectional method of Southern (1975). DNA immobilized on filters was then hybridized to nick-translated cDNA (Rigby *et al.*, 1977) and 5'-labeled bovine urothelial poly(A)$^+$ mRNA using standard conditions.

Subcloning and DNA sequencing

The 7.2-kb *Eco*RI and 1.5-kb *Hin*dIII fragments of clone λKB19[6] were purified by gel electrophoresis on low melting point agarose gels. After extraction, the DNA fragments were inserted into the appropriate sites of pUC18 (Norrander *et al.*, 1983). Restriction maps were constructed from the results of single and double restriction enzyme digests.

Nucleotide sequences of clones pKB19[1] and the two subclones of λKB19[6] were determined by the chemical degradation method (Maxam and Gilbert, 1980). Some fragments were also sequenced by the dideoxynucleotide sequencing method (Sanger *et al.*, 1977). Most of the sequence was determined in both directions. DNA sequence analysis, translation into amino acid sequences and comparisons with published sequences were performed with the aid of the program library for biological sequence analysis (for details see Rieger *et al.*, 1985; Magin *et al.*, 1986).

S1 nuclease mapping

Transcriptional mapping was essentially as described by Berk and Sharp (1977) using the modifications of Weaver and Weismann (1979). The DNA of a 1135-bp *Nco*I-fragment of a subclone (pKB40/E2) of λKB19[6] was isolated from a 1% low-melting agarose gel as described by Maniatis *et al.* (1982). The DNA fragment was 5' end-labeled, using calf intestinal alkaline phosphatase (Maniatis *et al.*, 1982), γ[^{32}P]ATP (Amersham, Bucks, UK) and T4 polynucleotide kinase (BRL, Bethesda) as described by Maxam and Gilbert (1980). This end-labeled fragment was digested with *Sma*I and the digest electrophoresed on a 1.5% low melting agarose gel. The DNA containing the 784-bp *Nco*I/*Sma*I fragment was isolated. Aliquots of the labeled fragments and 3 μg bovine urothelial poly(A)$^+$-mRNA were denatured at 80°C for 10 min and incubated for 16 h at 56, 58 and 60°C. After incubation, the samples were diluted with 250 μl of ice-cold 'S1-buffer' (0.25 M NaCl, 0.03 M sodium acetate, 1 mM $ZnSO_4$, pH 4.6) containing 125 units of S1 nuclease (Boehringer, Mannheim, FRG). The specifically hybridized RNA was further analyzed on a denaturing sequencing gel as described (Lehnert *et al.*, 1984).

Preparation of bovine cytokeratin 19 and digestion with chymotrypsin

A cytoskeletal fraction enriched in cytokeratin polypeptides was prepared from bovine bladder urothelium by extraction with buffers containing high concentrations of salt and Triton X-100 (Schiller *et al.*, 1982; Achtstaetter *et al.*, 1986). Cytokeratin polypeptides nos. 8 (A) and 19 (no. 22 of Schiller *et al.*, 1982) were purified by DEAE-anion exchange chromatography and reverse phase h.p.l.c. as described (Hatzfeld and Franke, 1985; Achtstaetter *et al.*, 1986). The purified lyophilized polypeptides were redissolved in 25 mM Tris-HCl buffer (pH 8) containing 9.5 M urea and 5% 2-mercaptoethanol. Equimolar amounts of cytokeratin polypeptides nos. 8 and 19 were mixed, diluted to a final concentration of 0.5 mg/ml and dialyzed stepwise to 2 mM Tris-HCl buffer (pH 8) and 20 mM Tris-

HCl buffer (pH 7.6) to induce IF formation which was controlled by electron microscopy (Hatzfeld and Franke, 1985). Limited proteolysis of the filaments thus formed with chymotrpysin (Sigma, St Louis, MO) was carried out at an enzyme:substrate ratio (w/w) of 1:250 for 75 min at room temperature. The reaction was stopped by applying the sample to an h.p.l.c.-column (BioRad RP304, BioRad Laboratories, Richmond, CA, USA) equilibrated in 35% vol acetonitrile that contained 0.07% trifluoroacetic acid (TFA, Fluka, Buchs, Switzerland) and 65% vol of 0.1% TFA. Peptide fragments were eluted with a gradient from 35 to 50% acetonitrile. Intact polypeptides as well as peptide fragments were checked for homogeneity by SDS–PAGE using 7–20% (w/v) gradient gels.

For sequencing, the peptides were submitted to automated Edman degradations using a gas-phase sequencer (Applied Biosystems, Forster City, CA) according to the method of Hunkapiller *et al.* (1983). Phenylthiohydantoin derivatives of amino acids were identified by h.p.l.c. (LKB; u.v. detector from Kratos Analytical, Ramsay, NJ) on a Lichrospher SuperRP-8 Column (Merck) using an isocratic elution buffer consisting of 68.5% (vol), 25 mM sodium acetate, 31.5% acetonitrile and 0.5% dichloroethane adjusted to pH 5.6.

Reconstitution of cytokeratin polypeptide no. 19

Purified lyophilized cytokeratin polypeptide no. 19 was redissolved in 25 mM Tris-HCl buffer (pH 8) containing 9.5 M urea, 5% 2-mercaptoethanol (ME) and a protein concentration of ~1 mg/ml. For stepwise reconstitution, the solution was dialyzed first to 4 M urea containing 5 mM Tris-HCl buffer (pH 8), then to 2 mM Tris-HCl buffer (pH 8), 10 mM Tris-HCl buffer (pH 7.5), 30 mM Tris-HCl buffer (pH 7.5), 50 mM Tris-HCl buffer (pH 7.5) and finally to 100 mM Tris-HCl buffer (pH 7.5), all buffers containing 1 mM dithiothreitol. Alternatively, 30 mM Tris-HCl buffer (pH 7.5) containing 50 or 100 mM NaCl was used. After each dialysis step, aliquots were taken and analyzed by electron microscopy using negative staining and the spraying-metal shadowing technique (Franke *et al.*, 1982a; Quinlan *et al.*, 1984).

Acknowledgements

We thank Drs P.Cowin and J.L.Jorcano for valuable discussions as well as for reading and correcting the manuscript and Drs G.Maier and H.Ponstingl (this institute) for the amino acid sequence determinations. The work has been supported by grants from the Deutsche Forschungsgemeinschaft (DFG) and the Ministry for Research and Technology of the Federal Republic of Germany.

References

Achtstaetter,T., Hatzfeld,M., Quinlan,R.A., Parmelee,D.C. and Franke,W.W. (1986) *Methods Enzymol.*, in press.

Ahmadi,B., Boston,N.M., Dobb,M.G. and Speakman,P.T. (1980) In Parry,D.A.D. and Creamer,L.K. (eds), *Fibrous Proteins: Scientific, Industrial and Medical Aspects.* Academic Press, New York, Vol. 2, pp. 161–166.

Bader,B.L. (1986) Diploma Thesis, Faculty of Biology, University of Heidelberg, pp. 1–156.

Balcarek,J.M. and Cowan,N.J. (1985) *Nucleic Acids Res.*, **13**, 5527–5543.

Banks-Schlegel,S.P. and Quintero,J. (1986) *Cancer Res.*, **46**, 250–258.

Bartek,J., Durban,E.M., Hallowes,R.C. and Taylor-Papadimitriou,J. (1985) *J. Cell Sci.*, **75**, 17–33.

Berk,A.J. and Sharp,P.A. (1977) *Cell*, **12**, 721–732.

Blake,C. (1983) *Trends Biochem. Sci.*, **8**, 11–13.

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

Chou,P.Y. and Fasman,G.D. (1978) *Annu. Rev. Biochem.*, **47**, 251–276.

Cooper,A.D. and Crain,W.R.,Jr. (1982) *Nucleic Acids Res.*, **10**, 4081–4092.

Craik,C.S., Rutter,W.J. and Fletterick,R. (1983) *Science*, **220**, 1125–1129.

Crewther,W.G., Dowling,L.M., Steinert,P.M. and Parry,D.A.D. (1983) *Int. J. Biol. Macromol.*, **5**, 267–274.

Doolittle,W.F.. (1978) *Nature*, **272**, 581–582.

Dynan,W.S. and Tjian,R. (1985) *Nature*, **316**, 774–778.

Eckert,R.L. and Green,H. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 4321–4325.

Fornwald,J.A., Kuncio,G., Peng,I. and Ordahl,C.P. (1982) *Nucleic Acids Res.*, **10**, 3861–3876.

Franke,W.W., Schmid,E., Osborn,M. and Weber,K. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 5034–5038.

Franke,W.W., Schiller,D.L., Moll,R., Winter,S., Schmid,E., Engelbrecht,I., Denk,H., Krepler,R. and Platzer,D. (1981a) *J. Mol. Biol.*, **153**, 933–959.

Franke,W.W., Winter,S., Grund,C., Schmid,E., Schiller,D.L. and Jarasch,E.D. (1981b) *J. Cell Biol.*, **90**, 116–127.

Franke,W.W., Schiller,D.L. and Grund,C. (1982a) *Biol. Cell*, **46**, 257–268.

Franke,W.W., Schmid,E., Schiller,D.L., Winter,S., Jarasch,E.-D., Moll,R., Denk,H., Jackson,B.W. and Illmensee,K. (1982b) *Cold Spring Harbor Symp. Quant. Biol.*, **46**, 431–453.

Franke,W.W., Schiller,D.L., Hatzfeld,M. and Winter,S. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 7113–7117.

Franz,J.K., Gall,L., Williams,M.A., Picheral,B. and Franke,W.W. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 6254–6258.

Fuchs,E. and Green,H. (1981) *Cell*, **25**, 617–625.

Fuchs,E.V., Coppock,S.M., Green,H. and Cleveland,D.W. (1981) *Cell*, **27**, 75–84.

Fuchs,E., Hanukoglu,I., Marchuk,D., Grace,M.P. and Kim,K.H. (1985) In Wang,E., Fischman,D., Liem,R.K.H. and Sun,T.-T. (eds), *Ann. New York Acad. Sci.*, **455**, 436–450.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.

Geisler,N. and Weber,K. (1982) *EMBO J.*, **1**, 1649–1656.

Geisler,N., Kaufmann,E. and Weber,K. (1982) *Cell*, **30**, 277–286.

Gilbert,W. (1978) *Nature*, **271**, 501.

Gilbert,W. (1985) *Science*, **228**, 823–824.

Gwo-Shu Lee,M., Lewis,S.A., Wilde,C.D. and Cowan,N.J. (1983) *Cell*, **33**, 477–487.

Hanukoglu,I. and Fuchs,E. (1982) *Cell*, **31**, 243–252.

Hanukoglu,I. and Fuchs,E. (1983) *Cell*, **33**, 915–924.

Hatzfeld,M. and Franke,W.W. (1985) *J. Cell Biol.*, **101**, 1826–1841.

Hoffmann,W. and Franz,J.K. (1984) *EMBO J.*, **3**, 1301–1306.

Hunkapiller,M.W., Hewick,R.M., Dryer,W.J. and Hood,L.E. (1983) *Methods Enzymol.*, **91**, 399–413.

Jonas,E., Sargent,T.D. and Dawid,I.B. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 5413–5417.

Johnson,L.D., Idler,W.W., Zhou,X.-M., Roop,D.R. and Steinert,P.M. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1896–1900.

Jorcano,J.L., Franz,J.K. and Franke,W.W. (1984a) *Differentiation*, **28**, 155–163.

Jorcano,J.L., Magin,T.M. and Franke,W.W. (1984b) *J. Mol. Biol.*, **176**, 21–37.

Jorcano,J.L., Rieger,M., Franz,J.K., Schiller,D.L., Moll,R. and Franke,W.W. (1984c) *J. Mol. Biol.*, **179**, 257–281.

Kaufmann,E., Weber,K. and Geisler,N. (1985) *J. Mol. Biol.*, **185**, 733–742.

Kim,K.H., Rheinwald,J.G. and Fuchs,E.V. (1983) *Mol. Cell. Biol.*, **3**, 495–502.

Kim,K.H., Schwartz,F. and Fuchs,E. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 4280–4284.

Krieg,T.M., Schafer,M.P., Cheng,C.K., Filpula,D., Flaherty,P., Steinert,P.M. and Roop,D.R. (1985) *J. Biol. Chem.*, **260**, 5867–5870.

Lazarides,E. (1982) *Annu. Rev. Biochem.*, **51**, 219–250.

Lehnert,M.E., Jorcano,J.L., Zentgraf,H., Blessing,M., Franz,J.K. and Franke,W.W. (1984) *EMBO J.*, **3**, 3279–3287.

Lonberg,N. and Gilbert,W. (1985) *Cell*, **40**, 81–90.

Lusky,M., Berg,L., Weiher,H. and Botchan,M. (1983) *Mol. Cell. Biol.*, **3**, 1108–1122.

Magin,T.M., Jorcano,J.L. and Franke,W.W. (1983) *EMBO J.*, **2**, 1387–1392.

Magin,T.M., Jorcano,J.L. and Franke,W.W. (1986) *Differentiation*, **30**, 254–264.

Maniatis,T., Frisch,E.F. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory Press, NY.

Marchuk,D., McCrohon,S. and Fuchs,E. (1984) *Cell*, **39**, 491–498.

Marchuk,D., McCrohon,S. and Fuchs,E. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1609–1613.

Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499–560.

McKeon,F.D., Kirschner,M.W. and Caput,D. (1986) *Nature*, **319**, 463–468.

McLachlan,A. (1978) *J. Mol. Biol.*, **98**, 293–304.

McLachlan,A.D. and Karn,J. (1982) *Nature*, **299**, 226–231.

Moll,R., Franke,W.W., Schiller,D.L., Geiger,B. and Krepler,R. (1982) *Cell*, **31**, 11–24.

Moll,R., Krepler,R. and Franke,W.W. (1983) *Differentiation*, **23**, 256–269.

Morris,A., Steinberg,M.L. and Defendi,V. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 8498–8502.

Mount,S.M. (1982) *Nucleic Acids Res.*, **10**, 459–472.

Naora,H. and Deacon,N.J. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 6196–6200.

Norrander,J., Kempe,T. and Messing,J. (1983) *Gene*, **26**, 101–106.

Parry,D.A.D., Steven,A.C. and Steinert,P.M. (1985) *Biochem. Biophys. Res. Commun.*, **127**, 1012–1018.

Perler,F., Efstradiadis,A., Lomedico,P., Gilbert,W., Kolodner,R. and Dodgson,J. (1980) *Cell*, **20**, 555–566.

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

Quax,W., van den Broek,L., Egberts,W.V., Ramaekers,F. and Bloemendal,H. (1985) *Cell*, **43**, 327–338.

Quinlan,R.A., Cohlberg,J.A., Schiller,D.L., Hatzfeld,M. and Franke,W.W. (1984) *J. Mol. Biol.*, **178**, 365–388.

Quinlan,R.A., Schiller,D.L., Hatzfeld,M., Achtstätter,T., Moll,R., Jorcano,J.L., Magin,T.M. and Franke,W.W. (1985) In Wang,E., Fischman,D., Liem,R.K.H. and Sun,T.-T. (eds) Intermediate Filaments. *Ann. New. York Acad. Sci.*, **455**, 282–306.

Rheinwald,J.G. and O'Connell,T.M. (1985) In Wang,E., Fischman,D., Liem,R.K.H. and Sun,T.-T. (eds), Intermediate filaments. *Ann. New York Acad. Sci.*, **455**, 259–267.
Rieger,M., Jorcano,J.L. and Franke,W.W. (1985) *EMBO J.*, **4**, 2261–2267.
Rigby,P.W.J., Dieckman,M., Rhodes,C. and Berg,P. (1977) *J. Mol. Biol.*, **113**, 237–251.
Romano,V., Hatzfeld,M., Magin,T.M., Zimbelmann,R., Franke,W.W., Maier,G. and Ponstingl,H. (1986) *Differentiation*, **30**, 244–253.
Ruppert,S., Scherer,G. and Schütz,G. (1984) *Nature*, **308**, 554–557.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Schiller,D.L., Franke,W.W. and Geiger,B. (1982) *EMBO J.*, **1**, 761–769.
Schmidt,W.N., Pardue,R.L., Tutt,M.C., Briggs,R.C., Brinkley,B.R. and Hnilica,L.S. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 3138–3142.
Singer,P.A., Trevor,K. and Oshima,R.G. (1986) *J. Biol. Chem.*, **261**, 538–547.
Southern,E. (1975) *J. Mol. Biol.*, **98**, 503–517.
Stone,E.M., Rothblum,K.N. and Schwartz,R.J. (1985) *Nature*, **313**, 498–500.
Steinert,P.M., Idler,W.W. and Goldman,R.D. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 4534–4538.
Steinert,P.M., Rice,R.H., Roop,D.R., Trus,B.L. and Steven,A.C. (1983) *Nature*, **302**, 794–800.
Steinert,P.M., Parry,D.A.D., Idler,W.W., Johnson,L.D., Steven,A.C. and Roop,D.R. (1985a) *J. Biol. Chem.*, **260**, 7142–7149.
Steinert,P.M., Steven,A.C. and Roop,D.R. (1985b) *Cell*, **42**, 411–419.
Strehler,E.E., Mahdavi,V., Periasamy,M. and Nadal-Ginard,B. (1985) *J. Biol. Chem.*, **260**, 468–471.
Südhof,T.C., Goldstein,J.L., Brown,M.S. and Russell,D.W. (1985) *Science*, **228**, 815–822.
Sun,T.-T. and Green,H. (1978) *Cell*, **14**, 469–476.
Sun,T.-T., Eichner,R., Schermer,A., Cooper,D., Nelson,W.G. and Weiss,R.A. (1984) In Levine,A.J., Vande Woude,G.F., Topp,W.C. and Watson,J.D. (eds), *Cancer Cells 1, The Transformed Phenotype*. Cold Spring Harbor Laboratory Press, NY, pp. 169–176.
Sun,T.-T., Tseng,S.C.G., Huang,A.J.-W., Cooper,D., Schermer,A., Lynch,M.H., Weiss,R. and Eichner,R. (1985) In Wang,E., Fischman,D., Liem,R.K.H. and Sun,T.-T. (eds), Intermediate filaments. *Ann. New York Acad. Sci.*, **455**, 307–329.
Tseng,S.C.G., Jarvinen,M., Nelson,W.G., Huang,J.-W., Woodcock-Mitchell,J. and Sun,T.-T. (1982) *Cell*, **30**, 361–372.
Tyner,A.L., Eichman,M.J. and Fuchs,E. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4683–4687.
Ueyama,H., Hamada,H., Battula,N. and Kakunaga,T. (1984) *Mol. Cell. Biol.*, **4**, 1073–1078.
Vasseur,M., Duprey,P., Brûlet,P. and Jacob,F. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1155–1159.
Weaver,R.F. and Weissmann,C. (1979) *Nucleic Acids Res.*, **7**, 1175–1193.
Weber,K. and Geisler,N. (1984) In Levine,A.J., Vande Woude,G.F., Topp,W.C. and Watson,J.D. (eds), *Cancer Cells 1, The Transformed Phenotype*. Cold Spring Harbor Laboratory Press, NY, pp. 153–159.
Weber,K. and Geisler,N. (1985) In Wang,E., Fischman,D., Liem,R.K.H. and Sun,T.-T. (eds), Intermediate filaments. *Ann. New York Acad. Sci.*, **455**, 126–143.
Woods,E.F. (1983) *Biochem. Int.*, **7**, 769–774.
Wu,Y.-J. and Rheinwald,J.G. (1981) *Cell*, **25**, 627–635.
Wu,Y.-J., Parker,L.M., Binder,N.E., Beckett,M.A., Sinard,J.H., Griffiths,C.T. and Rheinwald,J.G. (1982) *Cell*, **31**, 693–703.

Received on 29 April 1986