STRUCTURAL DIVERSITY AND EVOLUTION OF INTERMEDIATE FILAMENT PROTEINS[1]

Israel Hanukoglu

Department of Biology
Technion-Israel Institute of Technology
Haifa 32000, Israel

Elaine Fuchs

Department of Molecular Genetics and Cell Biology
The University of Chicago
Chicago, Illinois 60637 USA

ABSTRACT

The cytoskeletal network of most mammalian cells includes three major types of filamentous systems: microfilaments, intermediate filaments (IF) and microtubules with respective diameters of 6nm, 8-10nm and 25nm. Each of these filamentous systems is assembled from only one or two different subunits. The proteins that form these filaments are all encoded by multigene families, the members of which are differentially expressed in different tissues. While the protein sequences of actins and tubulins (which form microfilaments and microtubules, respectively) are highly conserved, the IF proteins show a

[1] Our work reviewed here was supported by a U.S. National Institutes of Health grant. I. H. was the recipient of a U.S. National Cancer Institute National Research Service Award. E. F. is the recipient of a National Institutes of Health Career Development Award and a Presidential Young Investigator Award.

much higher degree of diversity both in their sequences and the length of the polypeptide chains (Mr = 40-140K). Despite their diversity, the ultra-structures of the filaments formed by the different IF proteins are highly similar and these all resemble the structure of the microfibrils which form the backbone of such epidermal appendages as wool and hair. Analyses of the sequences of IF proteins and the microfibrillar α-keratins indicate that within the central 300 residues of these proteins there is a remarkable conservation of α-helical structural domains despite the divergence of sequences in this region. The IF proteins can be categorized into three major families on the basis of sequence homology in this central region: 1) type I keratins, 2) type II keratins, and 3) desmin, vimentin, glial filament protein and neurofilament proteins. The sequence homology within each group is > 50%, while between groups it is 24-35%. The terminal sequences on both sides of the central region are even more variable and the size heterogeneity among IF proteins is a result of the differences in the length of these terminal regions. The conservation of structural features despite diversity of the IF protein sequences and the juxtaposition of hyper-variable terminal sequences with relatively constant domains provide valuable data and pose intriguing questions about the mechanisms of evolutionary change.

I. INTRODUCTION

The cytoskeletal network of most mammalian cells includes three major types of filamentous systems: microfilaments, intermediate filaments (IF) and microtubules with respective diameters of 6nm, 8-10nm and 25nm. Each of these filamentous sytems is assembled from only one or two different subunits. The proteins that form these filaments are all encoded by multigene families, the members of which are differentially expressed in different tissues. While the protein sequences of actins and tubulins (which form microfilaments and microtubules, respectively) are highly conserved, the IF proteins show a much higher degree of diversity both in their sequences and the length of the polypeptide chains (M_r 40,000-140,000). Here, we shall present an

overview of our current knowledge of the sequence and structural relatedness of this most divergent group of cytoskeletal proteins and their genes.

II. MULTIPLICITY OF IF PROTEINS AND THEIR STRUCTURAL RELATIONSHIP

In mammalian organisms the total number of distinct proteins capable of forming IF in different tissues and at different stages of development may be at least 20-30 (Lazarides, 1982). The expression of subsets of this group of proteins is generally limited to specific types of tissues: keratins are expressed in epithelial cells, desmin in muscle cells, glial filament protein (GFP) in glial cells, neurofilament proteins (NFP) in neurons, and vimentin in mesenchymal cells and many other tissues. However, the IF proteins cannot all be grouped into strict categories based on their tissue specificity of expression as some of the IF proteins (especially vimentin together with others) coexist in different tissues, and the spectrum of IF proteins in a specific tissue may change during development and differentiation (Gard *et al.*, 1979; Franke *et al.*, 1982; Lazarides, 1982; Sharp *et al.*, 1982; Granger and Lazarides, 1983; Lane *et al.*, 1983; Ben-Zeev, 1984; Holthofer *et al.*, 1984). Proteins and genomic sequences that represent all of these five groups of IF proteins have been detected in all vertebrates examined to date (Fuchs and Marchuk, 1983; Quax *et al.*, 1984; Lewis *et al.*, 1984; Lewis and Cowan, 1985). There is immunological evidence that vimentin related polypeptides may also exist in invertebrates (Walter and Biessmann, 1984). Using cDNA probes, sequences homologous to the mRNAs of some IF proteins have been detected in organisms lower than the vertebrates; however, until the nature of the proteins coded by these sequences is established, these

remain as tentative indications for the presence of IF proteins in invertebrates (Fuchs and Marchuk, 1983).

Among the IF proteins, the keratins represent the largest and the most diverse group. The keratins which form IF are highly homologous to the keratins which form the backbone of such epidermal appendages as hair and wool. Thus, the multigene families of IF proteins also include these keratins. While in epithelial cells keratin filaments are organized in seemingly irregular patterns of a cytoskeletal network (henceforth named as cytoskeletal keratins), in epidermal appendages such as hair and wool the keratins form very orderly arrays of microfibrils (henceforth named as microfibrillar keratins) embedded in a rigid matrix of other proteins (much like steel rods embedded in concrete!) (Jones, 1975; for reviews see Crewther *et al.*, 1965; Fraser *et al.*, 1972; Fuchs and Hanukoglu, 1986). The proteins that constitute the matrix in epidermal appendages are also called keratins. However, as these "matrix keratins" are completely different from the microfibrillar α-keratins, both in terms of their sizes (6-20K vs. 40-70 K) and their primary and secondary structures, we shall not cover studies on these keratins here.

In vertebrates the total number of different keratins varies between 2-20 across species (Moll *et al.*, 1982; Fuchs and Marchuk, 1983). Most, if not all, keratins (both cytoskeletal and microfibrillar) can be grouped into two distinct classes of sequences (Crewther *et al.*, 1978, 1980; Fuchs *et al.*, 1981; Hanukoglu and Fuchs, 1982, 1983; Dowling *et al.*, 1983; Fuchs and Marchuk, 1983; Steinert *et al.*, 1983, 1984). These two classes were named as type I and type II keratins, respectively (Hanukoglu and Fuchs, 1983) extending a nomenclature used for wool keratin fragments (Crewther *et al.*, 1978). The type I keratins are relatively acidic (isoelectric pH 4.5-5.5) and small (M_r 40,000-59,000), whereas the type II keratins are more basic (isoelectric pH 6.5-7.5) and larger (M_r

53,000-67,000). In contrast to the other IF proteins, a single keratin polypeptide cannot polymerize into filaments by itself. The formation of keratin filaments requires a pairwise association and polymerization of at least two different keratins (Lee and Baden, 1976; Steinert *et al.*, 1976, 1982). *In vitro* polymerization studies indicate that the type I and type II keratins may be these necessary building blocks of keratin filaments (Franke *et al.*, 1983; Quinlan *et al.*, 1984). This hypothesis is further strengthened by the observations that at least one member of each of these two classes of keratins is present in all epithelial cells (Kim *et al.* 1983; Eichner *et al.*, 1984), and that type I and type II keratins and their corresponding genes are found in all vertebrates (Fuchs *et al.*, 1981; Fuchs and Marchuk, 1983).

The category of neurofilament proteins includes three distinct polypeptides (M_r 62,000, 107,000, and ~140,000) named, respectively, as NFP-L, NFP-M, and NFP-H (Geisler *et al.*, 1984, 1985). These three NFP, and the other non-keratin IF proteins, desmin, vimentin, and GFP represent single proteins and appear to be encoded by genes that are present in a single copy in the genome of all vertebrate species examined to date (Capetanaki *et al.*, 1983; Lewis and Cowan, 1985; Quax *et al.*, 1984; Zehner and Paterson, 1983; Lewis *et al.*, 1984). Consistent with the singular presence of desmin, vimentin, and GFP in most tissues, these proteins can assemble into IF by self-polymerization (homopolymer formation) (Steinert *et al.*, 1982; Geisler, Kaufmann and Weber, 1982). Desmin, vimentin and GFP can also form heteropolymer IF by copolymerization *in vitro* (Steinert *et al.*, 1982), as well as *in situ* in tissues and cells where these proteins coexist (Quinlan and Franke, 1982; Wang *et al.*, 1984). NFP-L can form filaments by itself but NFP-M, and NFP-H also participate in the formation of neurofilaments *in situ*

(Liem and Hutchison, 1982; Sharp *et al.*, 1982; Steinert *et al.*, 1982; Hirokawa *et al.*, 1984).

Despite major differences among the various IF proteins (especially the M_r range: 40,000-70,000, excluding NFP-M and NFP-H), the ultrastructures of the filaments formed by these proteins are highly similar (Kallman and Wessells, 1967; Henderson *et al.*, 1982; Milam and Erickson, 1982). Prior to the determination of the sequences of IF proteins and microfibrillar keratins, a number of biochemical, immunological and physiochemical studies indicated that both of these two groups of proteins are structurally related, and that both contain long regions of mainly α-helical conformation in the center, and a staggered conformation of unknown structure at the amino and carboxy terminal ends (Crewther and Harrap, 1967; Fraser *et al.*, 1972, 1976; Skerrow *et al.*, 1973; Jones, 1975; Steinert, 1978; Steinert, Idler and Goldman, 1980; Weber, Osborn and Franke, 1980). Since the early studies of Pauling and Corey (1953) and Crick (1953) it is thought that the helical regions of these proteins intertwine around one another to form a coiled-coil rod and that these rods are linked end-to-end to form a rope-like filament that constitutes a protofibril. A bundle of about 10 such protofibrils constitutes the microfibrils and IF. Although some early studies suggested that in each protofibril the number of polypeptide chains aligned side-by-side is three (Crewther and Harrap, 1967; Skerrow *et al.*, 1973; Steinert, 1978; Steinert, Idler and Goldman, 1980), more recent studies indicate that this number is two (Gruen and Woods, 1981; Geisler and Weber, 1982; Woods and Gruen, 1983; Quinlan *et al.*, 1984). Model building studies and analyses of the sequences of microfibrillar keratins also indicate that the basic unit of protofibrillar structure is made up of two intertwined α-helical polypeptide chains (Parry *et al.*, 1977; McLachlan, 1978).

The positioning of the non-helical terminal regions of the IF proteins in the overall structure of the protofibril is not known yet. The terminal sequences of IF proteins can be extensively digested without disrupting the ability of these proteins to assemble into filaments (Steinert *et al.*, 1983; Lu and Johnson, 1983). Yet, the helical segments without the terminal extensions cannot assemble into filaments by themselves (Geisler *et al.*, 1982). Thus, it is currently thought that the terminal domains largely project from the central coiled-coil rod and contribute to both end-to-end linkage of IF proteins within the protofibril and the interactions of these proteins with other cellular molecules.

III. COMPARISON OF THE SEQUENCES AND PREDICTED SECONDARY STRUCTURES OF THE IF PROTEINS

Currently, the complete or nearly complete sequences of over 10 different IF proteins and microfibrillar keratins are known (Fig. 1). Analysis of these sequences indicate two major structural motifs: 1) A central region of about 300 residues predominantly in α-helical conformation, and 2) amino and carboxy terminal regions that show an amazing variability in their lengths and sequences across different IF proteins subclasses, but which appear to be conserved across species for each subclass. These results are in general consistent with the observations noted in the previous section. However, the analyses based primarily on sequence data present a more precise model for the structure of the IF proteins that is, in some important respects, different from those of some of the earlier studies. (For a detailed discussion see Geisler and Weber, 1982). A comparison of the sequences and the


Amino terminus
Helical domain I
Helical domain II
Helical domain III
Helical domain IV
Carbory terminus

Figure 1

Figure 1 Legend. The amino acid sequences of IF proteins and the common location of the predicted helical domains. The sequences are from the following sources: 1) MT-I: Type I cytoskeletal keratin from mouse epidermis (Mr = 59K) (Steinert *et al.*, 1983); 2) HT-I: Type I cytoskeletal keratin from human epidermal cells (Mr = 50K) (Hanukoglu and Fuchs, 1982; Marchuk *et al.*, 1984); 3) HT-II: Type II cytoskeletal keratin from human epidermal cells (Mr = 56K) (Hanukoglu and Fuchs, 1983); 4) MT-II: Type II cytoskeletal keratin from mouse epidermal cells (Mr = 60K) (Steinert *et al.*, 1984); 5) CDES: Desmin from chicken muscle (Geisler and Weber, 1982); 6) HVIM: Vimentin from hamster lens (Quax *et al.*, 1983); 7) MGFP: Glial fibrillary protein from mouse brain (Lewis *et al.*, 1984); 8) PNFP: Neurofilament protein-M from porcine spinal cord (Geisler *et al.*, 1984). Table I includes additional sequence comparisons with wool microfibrillar keratins: WT-I: Type I microfibrillar keratin from sheep wool (Crewther *et al.*, 1980; Gough *et al.*, 1978; Dowling *et al.*, 1983); WT-II: Type II microfibrillar keratin from sheep wool (Crewther *et al.*, 1978; 1980; Gough *et al.*, 1978; Sparrow and Inglis, 1980; Dowling *et al.*, 1983). The dots (.) indicate missing sequence information. In cases where the sequence of the amino terminus is not known the numbering on the left side of the sequences start with the first known amino terminal residue of the longest segment. The gaps in the sequences indicate gaps introduced in order to align the sequences for optimal homology. The position of the helical domains are based on computerized secondary structure prediction analyses using the Chou and Fasman (1978, 1979) and Garnier, Osguthorpe and Robson (1978) methods as previously described (Hanukoglu and Fuchs, 1982; 1983).

predicted secondary structures of these proteins indicate the following major similarities and differences among the IF proteins (for references see legend of Fig. 1).

A. The Central Helical Region Is Structurally Conserved

The central approximately 300 residue long portion of all IF protein sequences can be aligned for optimal homology without any or only a few (1-6) gaps. In this central region different IF proteins share 24% to 90% homology with each other (Fig. 1, Table I). The IF proteins can be

Table I. Percent homology between the central regions of IF proteins (see Fig. 1).

	HT-I	MT-I	WT-I	HT-II	MT-II	WT-II	CDES	HVIM	MGFP	PNFP
HT-I		66.0	56.8	26.6	25.9	29.7	35.2	34.6	37.1	35.0
MT-I			50.2	24.4	23.7	27.6	32.9	32.9	34.2	32.2
WT-I				26.2	24.7	28.0	30.8	30.5	33.3	32.3
HT-II					88.8	58.6	34.9	34.9	32.7	31.9
MT-II						54.9	34.4	34.4	32.8	32.1
WT-II							35.9	37.1	37.1	32.9
CDES								74.1	63.4	47.8
HVIM									63.1	50.0
MGFP										48.1
PNFP										

categorized into three major families on the basis of sequence homology in this central region: 1) type I keratins; 2) type II keratins; and 3) desmin, vimentin, GFP, and neurofilament proteins. The sequence homology within each group is > 50%, while between groups it is 24-35%.

The presently available few sequences of type I and type II keratin suggest that within both of these groups cytoskeletal and microfibrillar keratins constitute distinct subfamilies with a higher degree of homology by at least 10% (Table I). However, it remains to be seen whether this homology difference will hold when more sequences from both putative subfamilies will be known.

Even though there is only a low degree of homology between some of the IF protein sequences, the secondary structures of the central regions of all IF proteins seem to be remarkably conserved. Secondary structure prediction analyses using the methods of Chou and Fasman (1978, 1979) and Garnier *et al.* (1978) indicate that within this central region, there are four richly α-helical domains which are demarcated from one another by three regions for which β-turns are predicted with a high degree of probability in all sequences. The first two of these β-turn regions

contain proline(s) in some, but not all sequences. The four helical domains (marked as I, II, III, and IV in Fig. 1) are predicted to be nearly constant in size in all IF proteins and they are approximately 30-40, 100, 35-40, and 100 residues long, respectively. Different secondary structure prediction analyses yield slightly different lengths for the helical domains (see references in Fig. 1 legend). Thus, the predictions of the points of initiation and ends of the helical domains are not precise, and these remain to be determined by physicochemical structural analyses.

Within the predicted helical domains both charged and hydrophobic residues display specific periodicities. The hydrophobic residues predominantly occupy positions **a** and **d** in consecutive heptade repeats of a-b-c-d-e-f-g. This periodicity is characteristic of coiled-coil α-helical super-secondary structures and thus provides further support for the validity of the predicted positions of the helical domains (Fraser *et al.*, 1976; Parry *et al.*, 1977; Elleman *et al.*, 1978; McLachlan, 1978; McLachlan, and Stewart, 1982; references in Fig. 1 legend). In these domains both the hydrophobic and charged residues are conserved more frequently and many substitutions for them represent conservative replacements, e.g., Asp (-) for Glu (-), or Arg (+) for Lys (+).

Although among all IF proteins amino acid sequence homology is higher within the predicted helical domains, it is especially prominent in domain III and in the carboxy terminal end of domain IV.

Despite evolutionary divergence of sequences, the amino acid compositions of the total central region and those of the individual helical domains of IF proteins have remained remarkably similar. The conservation of the helical secondary structure of IF proteins, despite the divergence of their sequences, may be partly a result of this conservation of amino acid compositions compatible with α-helicity. For example, for each IF protein 25-30% of the residues in this region are Glu and Leu, both

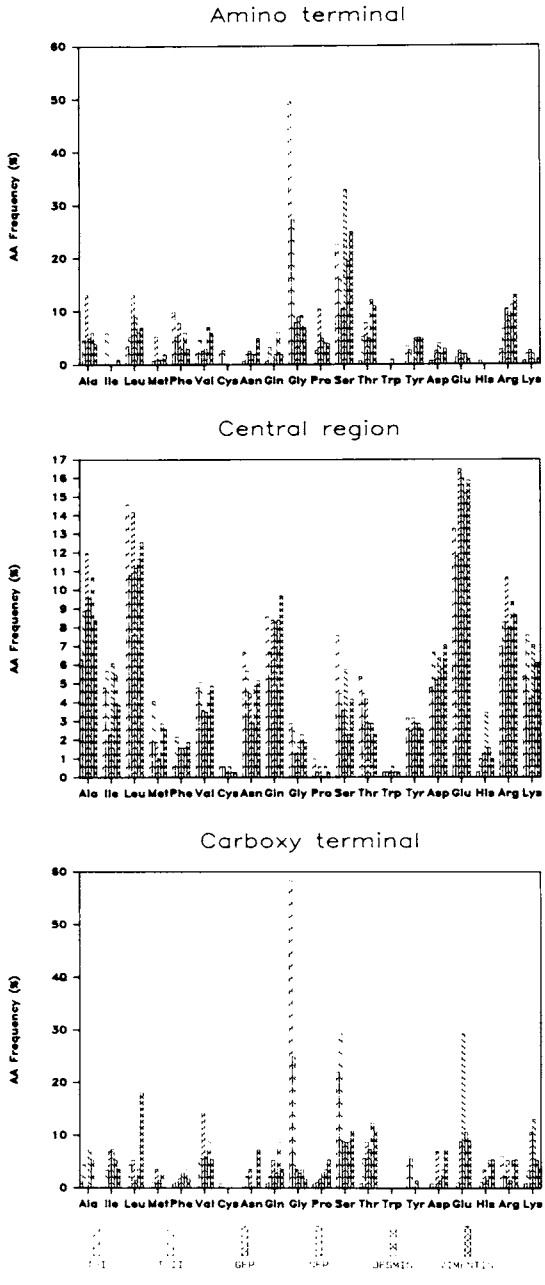
Amino terminal
AA Frequency (%)
Ala Ile Leu Met Phe Val Cys Asn Gln Gly Pro Ser Thr Trp Tyr Asp Glu His Arg Lys
Central region
AA Frequency (%)
Ala Ile Leu Met Phe Val Cys Asn Gln Gly Pro Ser Thr Trp Tyr Asp Glu His Arg Lys
Carboxy terminal
AA Frequency (%)
Ala Ile Leu Met Phe Val Cys Asn Gln Gly Pro Ser Thr Trp Tyr Asp Glu His Arg Lys

Figure 2

Figure 2 Legend. The amino acid composition of the amino terminal, central, and carboxy terminal regions of IF proteins. The IF protein sequences included in this figure are shown in Figure 1. The keratins, marked as T-I and T-II, are mouse keratins. The "central region" extends from the first residue of the first helical domain to the last residue of the fourth helical domain. The terminal regions are as shown in Figure 1.

of which highly favor α-helical structures (Fig. 2). It should be noted that the significant homologies among IF proteins do not result from this conserved amino acid composition, as a shift of the optimal alignment of the sequences by a single residue reduces the homologies to random levels.

B. The Amino and Carboxy Terminal Regions are Extremely Variable Across Subclasses of IF Proteins

The amino and carboxy terminal regions of different IF proteins show an extreme degree of heterogeneity in their sizes, sequences, and amino acid compositions (Figs. 1 and 2). The size differences among IF proteins is a result of differences in the length of these terminal portions rather than the structurally conserved central region. In terms of amino acid composition, the only conspicuous feature common to *all* IF proteins appears to be a relatively high ratio of serines in the amino and, to a lesser degree, carboxy terminal regions (Fig. 2). The secondary and higher orders of structure of these regions are not known yet, and structure prediction analyses cannot be applied to them because of their unusual sequences and highly skewed amino acid compositions. Thus, future structural information may disclose novel features at the ends of IF proteins.

Although the terminal regions show a great diversity, the members of each of the three major families (type I keratins, type II keratins, and

non-keratins) can be further subgrouped on the basis of sequence homology of these terminal regions as noted below. The available sequences of IF proteins from different species indicate that for each subclass of IF protein the sequences of these regions are conserved across species. However, the extent of the evolutionary conservation of terminal region sequences of each subclass remains to be further defined by determination of sequences from a greater number of species.

1. Type I and type II keratins

a) *Within both type I and type II keratin families, the cytoskeletal and microfibrillar keratins can be distinguished by their terminal sequences.* Although the central helical domains of the cytoskeletal and microfibrillar keratins share 50-60% homology within each type, the known terminal sequences of these two groups of proteins show no homology (Fuchs and Hanukoglu, 1985). In the terminal regions both wool microfibrillar type I and type II keratins are rich in cysteine and proline (Crewther *et al.*, 1980; Sparrow and Inglis, 1980). The cysteine richness of these regions readily suggests that these residues are involved in S-S covalent bond formation between individual keratins contributing greatly to the special characteristics and strength of the microfibrils. In contrast, the terminal regions of cytoskeletal type I and type II keratins contain no or only a few cysteines or prolines, but instead they are highly rich in glycine as noted below. However, as the sequences of only a few members of these subfamilies are known, their characteristics cannot be outlined fully yet. In addition, the possibility of the existence of keratins that share properties of both forms also remains.

b) *The cytoskeletal keratins can be further subgrouped on the basis of evolutionarily conserved differences in the terminal sequences.* The amino acid compositions of both cytoskeletal type I and

type II keratins are unusually rich in glycine. The great majority of these glycines appear in the terminal regions of the keratins as glycine is not compatible with α-helix. The amino termini of all cytoskeletal keratin sequences known to date include irregularly spaced short repeats of three or four glycines separated by other residues or clusters of serines, whereas the carboxy termini of some but not all cytoskeletal keratins include GGGX type repeats (Fig. 1).

At present, the partial sequences of only two different type I keratins from one species are known (Jorcano *et al.*, 1984). The central helical regions of these bovine keratins are highly homologous with one another (68% homology) as well as with those of other type I keratins from human, mouse and Xenopus (Hanukoglu and Fuchs, 1982; Steinert *et al.*, 1983; Hoffman and Franz, 1984). However, the carboxy termini of these two bovine keratins do not show even a vestige of sequence homology. On the basis of a comparison of these two sequences with the sequences of type I keratins from other species, one subfamily of type I keratins may be characterized by a carboxy terminus that includes GGGXYGG type repeats, whereas the other by a non-glycine rich carboxy terminus with a specific conserved sequence. Most interestingly, the 3' non-coding regions of the mRNAs of these subfamilies are also conserved (Jorcano *et al.*, 1984). Most probably, as more keratin sequences from different species are determined, other similar evolutionarily conserved distinct variations will be observed in both type I and type II keratin families.

As noted above, in contrast to non-keratins, the assembly of a keratin filament requires the pairwise association of a type I keratin with a type II keratin. The microfibrillar keratins include specific structural pairs of type I and type II keratins. Whether similarly some or all of the cytoskeletal type I keratin subfamilies structurally match specific type II

keratin subfamilies remains to be determined. This subject can be examined from two perspectives: Are specific type I and type II subfamilies expressed together in different tissues? Which specific type I and type II subfamilies can copolyermize *in vitro*?

2. Non-keratins

a) *Desmin, vimentin and GFP which can form both homopolymers and heteropolymers with one another, share similar terminal sequences.* Although desmin, vimentin and GFP share a lower homology with one another in their terminal sequences than in their central regions, these terminal regions are much less divergent than those of the keratin subfamilies (Quax *et al.*, 1984). As the terminal sequences of the IF proteins appear to be involved in end-to-end linkage of protofibrillar subunits, the similarity of these sequences may partly explain the ability of these proteins to copolymerize both *in vitro* and *in situ* (Steinert *et al.*, 1982; Wang *et al.*, 1984).

b) *Among IF proteins the most extreme size variation is observed in the terminal regions of NFP.* As noted above, in vertebrates the NF are constituted from three different proteins NFP-L, NFP-M, and NFP-H, with respective M_r of 62,000, 107,000 and ~140,000. Despite this extreme size variation, all three NFP share the same α-helical central region common to all IF proteins, and the size differences result from the different sizes of the carboxy terminal regions (Geisler *et al.*, 1984, 1985). The partial sequences and amino acid compositions of fragments isolated from the terminal regions indicate that they are highly variable in sequence but generally share a high percentage of glutamate (20-30%) and lysine (16-25%) (Geisler *et al.*, 1984, 1985).

IV. COMPARISON OF THE GENE STRUCTURES OF IF PROTEINS

Recently, the complete sequences of a hamster vimentin and a human type I keratin genes have been determined (Quax *et al.*, 1983; Marchuk *et al.*, 1984, 1985). These two sequences as well as mapping of intron positions in bovine type I and type II keratin genes (Lehnert *et al.*, 1984) reveal the following major similarities and differences among IF protein genes:

The positions of introns are remarkably well conserved among all IF protein genes. Both human and bovine type I keratin genes characterized to date have seven introns, while bovine type II keratin genes and the hamster vimentin gene have eight introns. In each gene all but one of the introns appear in the central helical region of the coded protein. Six of these introns occupy identical positions in this structurally conserved region.

In many genes intron positions have been observed to mark the boundaries of structural or functional domains of the coded proteins (Craik, 1983; Go, 1983; Lonberg and Gilbert, 1985). However, in IF protein genes most of the highly conserved intron positions do not appear to border the structural domains of the proteins. Only one intron (intron 6 in type I keratin and vimentin, and intron 7 in type II keratin) occurs at the end of the central helical region. The conserved presence of this intron at the border of the conserved central region and the highly variable carboxy terminal region raises the possibility that intron mediated gene fragment shuffling might have resulted in the juxtaposition of a structurally conserved region with highly variable segments. However, the other introns in the central region are all located within the helical domains. As noted above, both the sequence and the structure of the helical domains are conserved much more than other regions of IF proteins. Introns appear

neither at the border of the first helical domain and the highly divergent amino terminal region, nor at the borders of the helical domains and the proline and glycine rich non-helical linker regions that connect these domains (Fig. 1; Quax *et al.*, 1983; Marchuk *et al.*, 1984). The conservation of the positions of the introns suggests that these may be of some functional significance. Yet, with our present knowledge of the structure of the IF proteins if these positions have a structural significance it eludes us.

Despite the strict conservation of intron position, across and within species comparisons of the intron sequences indicate that neither the sizes nor the sequences of introns are conserved among IF protein genes.

The regulatory sequences 5' upstream from the type I keratin and the vimentin genes appear to be very divergent. Furthermore, the type I keratin gene, which is highly expressed in some epidermal tissues, contains sequences that share significant homology with enhancer elements found in other genes, whereas vimentin gene does not appear to have similar sequences in the corresponding gene segments. These differences in the 5' regulatory regions may ultimately explain the differential regulation of expression of IF genes in different tissues (Marchuk *et al.*, 1985).

V. SEQUENCE BASED CLASSIFICATION OF IF PROTEINS

The IF proteins have been sometimes referred to as a "family of proteins." Yet, as detailed above, sequence homology between *some* IF proteins can be about 10% over the complete lengths of the proteins and as low as 24% in the central structurally conserved region (Table I). This degree of diversity of sequences mandates the use of a larger set name for the IF proteins. Thus, consistent with nomenclature used for other

multigene families (e.g., Regier *et al.*,1983), we suggest that the group of IF proteins be referred to as a "*superfamily*".

To summarize the sequence comparisons presented above, the 20-30 proteins which constitute the IF protein superfamily can be grouped in three distinct families based on their sequence homologies in the structurally conserved central region: 1) type I keratins, 2) type II keratins, 3) non-keratins. The first two families include about 10 members each in mammals, whereas the third includes desmin, vimentin, GFP, NFP-L, NFP-M, and NFP-H. The degree of central region homology between proteins from different families is 24-35% and between proteins within the same family it is > 50% (Table I). The proteins in each of the three families, but especially the multi-membered keratin families, can be further grouped into subfamilies primarily by the similarities and differences of the sequences of their amino and carboxy terminal regions.

This classification of IF proteins based strictly on sequence similarities also reflects developmental, structural, and functional differences among the proteins. The keratins are expressed in epithelial tissues and apparently can form filaments only by association of specific type I and type II pairs, whereas the non-keratins are expressed predominantly in non-epithelial tissues and can form filaments by both self-polymerization and, in tissues and cells where they co-exist, by hetero-polymer formation in association with another non-keratin IF protein.

VI. THE EVOLUTION OF IF PROTEINS

The characteristics common to all IF proteins and their genes provide strong evidence that all IF proteins have evolved from a common ancestral protein. Thus, we can surmise that the IF proteins that exist today have

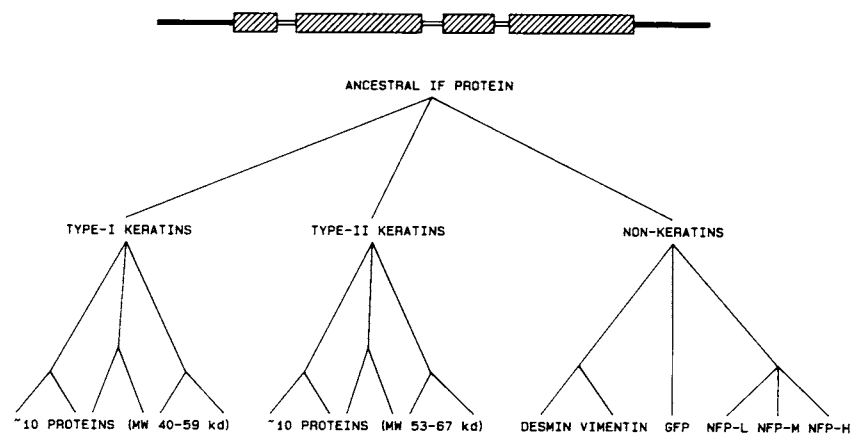

Figure 3. A hypothetical scheme for the evolution of the IF proteins. The top portion of the figure shows a secondary structural model for the ancestral IF protein based on Figure 1. In this model the bars with the parallel lines represent the four α-helical domains; the empty bars represent the short β-turn regions that link the α-helical domains; and the black bars represent the non-helical terminal regions that highly differ in length and sequence among the different IF proteins (Fig. 1). The lengths of the domains within the central region are depicted in proportion to the lengths of the respective domains (Fig. 1). The nodes from which three lines emanate indicate uncertainty in the respective order of emergence of the different IF protein classes during evolution. The branches of keratin subfamilies are hypothetical and as only a few sequences of keratins are known the number of subfamilies cannot be estimated at present.

probably evolved from an ancestral protein which also possessed a predominantly α-helical central region (Fig. 3). Yet the nature of the ancestral terminal sequences on either side of this helical region cannot be conjectured because of the extreme variability of these in present day IF proteins. The possibility that the ancestral protein had no, or very

short, terminal sequences also exists, as one cytoskeletal keratin has an M_r of 40,000, which indicates that its terminal sequences are very short (Fuchs and Marchuk, 1983; Kim *et al.*, 1984).

The groupings of IF proteins by central and terminal region sequence homologies indicate that the evolution of IF proteins probably proceeded in two major stages: 1) Generation of the prototypes of the three major families; 2) The evolution of subfamilies of proteins from these prototypes mainly by diversification of the sequences of the terminal regions, and to a minor degree by divergence of the sequence of the central region (Fig. 3). The evolutionary paths that have led to both of these stages most probably consisted of a series of events that included gene duplications, and subsequent point mutations, insertions and deletions, as previously suggested for other multigene families (e.g., Markert *et al.*, 1975; Wahli *et al.*, 1981; Jones and Kafatos, 1982).

The great differences in the types of evolutionary modifications observed in the central versus terminal regions of IF proteins represent outstanding examples to illustrate the interdependence of evolutionary changes and structural and functional roles of protein segments. While, during the course of evolution, very few insertions and deletions have been accepted in the central region, the terminal regions have been completely redesigned and structured for different IF proteins.

The conservation of the size and the secondary structural organization of the central region during evolution was probably necessary to maintain the ability of these proteins to assemble into coiled-coil filaments. Between some IF proteins the overall sequence of this region shows up to 76% divergence. Hence, structural conservation cannot be ascribed solely to the maintenance of a similar sequence, but also to the maintenance of (1) an α-helix compatible amino acid composition, and (2) the specific periodicities of charged and hydrophobic residues which

delineate points of interaction between helices of polypeptide chains forming coiled-coils (see Section III.A)

The high degree of divergence of the sequences of the central regions indicates that many details of the surface topology of the IF proteins (the R groups of amino acid residues project outwards from the main axis of an α-helix) are not crucial for defining the similarity of their secondary and super-secondary structures. This idea is strengthened further by the observation that IF can be formed *in vitro* and *in situ* by combinations of several different IF proteins. This degree of freedom in the choice of amino acids that make up the sequences of IF proteins greatly contrasts with the strict conservation of the sequences of globular proteins that form cytoskeletal filaments -- actin and tubulin. For example, between human and *Drosophila* cytoplasmic actin sequences, there is only a 2% divergence (Hanukoglu *et al.*, 1983). This difference suggests that the globular shapes of actins and tubulins may be disrupted more readily by single residue alterations and that their surface topology is much more precisely architectured for their many structural interactions, whereas the long helical segments of IF proteins appear to tolerate substitutions easily as long as specific residues at points of interaction between different helical domains are maintained.

The amino and carboxy terminal regions of IF proteins are extremely diversified across subfamilies, but the presently available few sequences from different species suggest that the representative of each subfamily is conserved across species (see Section III.B.1 and III.B.2). If these observations are generalizable to a larger number of species, then the across-species conservation of the unusual terminal sequences would suggest that they fulfill indispensable roles in the structure of IF proteins within the protofibril, and the interactions of these proteins with other cellular molecules. Thus, the different terminal sequences may be

important in selective interactions with specific molecules in different tissues. Yet, the question that remains is to what degree the complete sequences of the terminal regions are necessary for these specific interactions? The possibility can be raised that these terminal regions include segments without structural importance on the basis of the finding that partial digestion of these do not seem to disrupt the ability of IF proteins to form IF (Section II.B). But, in the absence of good understanding of the structural organization and interactions of these regions and the degree of their across-species conservation for each subfamily, these possibilities remain speculations and the posed question cannot be answered reliably at present.

Understanding the mechanisms and selective pressures that have led to the generation and diversification of the unusual amino and carboxy terminal sequences pose some of the biggest puzzles in the evolution of IF proteins. The following questions represent some of these: How were the extremely variable amino and carboxy terminal regions juxtaposed with the conserved central region? How did the glycine rich inexact repeats at the amino and carboxy terminal regions of cytoskeletal keratins arise? Do they share a common origin? Judging from their conserved central region sequences, the microfibrillar and cytoskeletal keratin subfamilies are much more closely related within each type I and type II family. But, while the terminal sequences of these subfamilies appear to be completely unrelated, cytoskeletal keratins from different families show similar glycine rich repeats. The sequences of these terminal regions cannot be aligned with a degree of certainty, but do their glycine rich repeats nonetheless indicate a common origin, or represent a case of convergent evolution serving the need of these proteins to assemble together? How did the carboxy terminal portions of NFP-M and NFP-H reach their gigantic sizes and sequences different from all other IF proteins?

In conclusion, the many IF proteins with their segmented architecture of a conserved central α-helical region flanked by extremely variable amino and carboxy terminal sequences provide valuable data and pose intriguing questions about the mechanisms of evolutionary change.

REFERENCES

Ben-Zeev, A. (1984). Control of intermediate filament protein synthesis by cell-cell interaction and cell configuration. *FEBS Lett.* **171**, 107-110.

Chou, P. Y., and Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* **47**, 45-148.

Chou, P. Y., and Fasman, G. D. (1979). Prediction of β-turns. *Biophys. J.* **16**, 367-383.

Craik, C. S., Rutter, W. J., and Fletterick, R. (1983). Splice junctions: Association with variation in protein structure. *Science* **220**, 1125-1129.

Capetanaki, Y. G., Ngai, J., Flytzanis, C. N., and Lazarides, E. (1983). Tissue-specific expression of two mRNA species transcribed from a single vimentin gene. *Cell* **35**, 411-420.

Crewther, W. G., and Harrap, B. S. (1967). The preparation and properties of a helix-rich fraction obtained by partial proteolysis of low sulfur S-carboxymethyl-kerateine from wool. *J. Biol. Chem.* **242**, 4310-4319.

Crewther, W. G., Inglis, A. S., and McKern, N. M. (1978). Amino acid sequences of α-helical segments from S-carboxymethylkerateine-A. *Biochem. J.* **173**, 365-371.

Crewther, W. G., Dowling, L. M., and Inglis, A. S. (1980). Amino acid sequence data from a microfibrillar protein of α-keratin. *In* "The Structure and Chemical Reactions of Keratins," Vol. 2, pp. 79-91.

Crick, F. H. C. (1953). The packing of α-helices: simple coiled-coils. *Acta Cryst.* **6**, 689-697.

Dowling, L. M., Parry, D. A. D., and Sparrow, L. G. (1983). Structural homology between hard α-keratin and the intermediate filament proteins desmin and vimentin. *Biosci. Rep.* **3**, 73-78.

Eichner, R., Bonitz, P., and Sun, T.-T. (1984). Classification of epidermal keratins according to their immunoreactivity, isoelectric point and mode of expression. *J. Cell Biol.* **98**, 1388-1396.

Elleman, T. C., Crewther, W. G., and Touw, J. V. D. (1978). Amino acid sequences of α-helical segments from S-carboxymethylkerateine-A: Statistical analysis. *Biochem. J.* **173**, 387-391.

Franke, W. W., Schmid, E., Schiller, D. L., Winter, S., Jarasch, E. D., Moll, R., Denk, H., Jackson, B. W., and Illmensee, K. (1982). Differentiation-related patterns of expression of protein of intermediate-size filament in tissues and cultured cells. *Cold Spring Harbor Symp. Quant. Biol.* **46**, 431-474.

Franke, W. W., Schiller, D. L., Hatzfeld, M., and Winter, S. (1983). Protein complexes of intermediate-sized filaments: melting of cytokeratin complexes in urea reveals different polypeptide separation characteristics. *Proc. Natl. Acad. Sci. USA* **80**, 7113-7117.

Fraser, R. D. B., MacRae, T. P., and Rogers, G. E. (1972). "Keratins: Their Composition, Structure and Biosynthesis. C. C. Thomas, Springfield, USA.

Fraser, R. D. B., MacRae, T. P., and Suzuki, E. (1976). Structure of the α-keratin microfibril. *J. Mol. Biol.* **108**, 435-452.

Fuchs, E. V., Coppock, S. M., Green, H., and Cleveland, D. W. (1982). Two distinct classes of keratin genes and their evolutionary significance. *Cell* **27**, 75-84.

Fuchs, E., and Marchuk, D. (1983). Type I and type II keratins have evolved from lower eukaryotes to form the epidermal intermediate filaments in mammalian skin. *Proc. Natl. Acad. Sci. USA* **80**, 5857-5861.

Fuchs, E., and Hanukoglu, I. (1986). Epidermal α-keratins: Structural diversity and changes during tissue differentiation. *In* "The Biology of the Integument" (in press).

Gard, D. L., Bell, P. B., and Lazarides, E. (1983). Coexistence of desmin and the fibroblastic intermediate filament subunit in muscle and nonmuscle cells: Identification and comparative peptide analysis. *Proc. Natl. Acad. Sci. USA* **76**, 3894-3898.

Garnier, J., Osguthorpe, D. J., and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97-120.

Geisler, N., and Weber, K. (1982). The amino acid sequence of chicken muscle desmin provides a common structural model for intermediate filament proteins. *EMBO J.* **1**, 1649-1656.

Geisler, N., Kaufmann, E., and Weber, K. (1982). Protein chemical characterization of three structurally distinct domains along the protofilament unit of desmin 10 nm filaments. *Cell* **30**, 277-286.

Geisler, N., Fischer, S., Vandekerckhove, J., Plessman, U., and Weber, K. (1984). Hybrid character of a large neurofilament protein (NF-M):

intermediate filament type sequence followed by a long and acidic carboxy-terminal extension. *EMBO J.* **4**, 57-63.

Geisler, N., Fischer, S., Vandekerckhove, J., Van Damme, J., Plessman, U., and Weber, K. (1985). Protein-chemical characterization of NF-H, the largest mammalian neurofilament component; intermediate filament-type sequences followed by a unique carboxy terminal extensions. *EMBO J.* **4**, 57-63.

Go, M. (1983). Modular structural units, exons, and function in chicken lysozyme. *Proc. Natl. Acad. Sci. USA* **80**, 1964-1968.

Gough, K. H., Inglis, A. S., and Crewther, W. G. (1978). Amino acid sequences of α-helical segments from S-carboxymethylkerateine-A. *Biochem. J.* **173**, 373-385.

Granger, B. L., and Lazarides, E. (1983). Expression of major neurofilament subunit in chicken erythrocytes. *Science* **221**, 553-556.

Gruen, L. C., and Woods, E. F. (1983). Structural studies on the microfibrillar proteins of wool. *Biochem. J.* **209**, 587-595.

Hanukoglu, I., and Fuchs, E. (1982). The cDNA sequence of a human epidermal keratin: divergence of sequence but conservation of structure among intermediate filament proteins. *Cell* **31**, 243-252.

Hanukoglu, I., and Fuchs, E. (1983). The cDNA sequence of a type II cytoskeletal keratin reveals constant and variable structural domains among keratins. *Cell* **33**, 915-924.

Hanukoglu, I., Tanese, N., and Fuchs, E. (1983). Complementary DNA sequence of a human cytoplasmic actin. *J. Mol. Biol.* **163**, 673-678.

Hirokawa, N., Glicksman, M. A., Willard, M. B. (1984). Organization of mammalian neurofilament polypeptides within the neuronal cytoskeleton. *J. Cell Biol.* **98**, 1523-1536.

Henderson, D., Geisler, N., and Weber, K. (1982). A periodic ultrastructure in intermediate filaments. *J. Mol. Biol.* **155**, 173-176.

Hoffman, W., and Franz, J. K. (1984). Amino acid sequence of the carboxy-terminal part of an acidic type I cytokeratin of molecular weight 51000 from Xenopus laevis epidermis as predicted from the cDNA sequence. *EMBO J.* **3**, 1301-1306.

Holthofer, H., Miettinen, A., Lehto, V.-P., Lehtonen, E., and Virtanen, I. (1984). Expression of vimentin and cytokeratin types of intermediate filament proteins in developing and adult human kidneys. *Lab. Invest.* **50**, 552-559.

Jones, C. W., and Kafatos, F. C. (1982). Accepted mutations in a gene family: Evolutionary diversification of duplicated DNA. *J. Mol. Evol.* **19**, 87-103.

Jones, L. N. (1975). The isolation and characterization of α-keratin microfibrils. *Biochim. Biophys. Acta.* **412**, 91-98.

Jorcano, J. L., Rieger, M., Franz, J. K., Schiller, D. L., Moll, R., and Franke, W. W. (1984). Identification of two types of keratin polypeptides within the acidic cytokeratin subfamily I. *J. Mol. Biol.* **179**, 257-281.

Kallman, F., and Wessells, N. K. (1967). Periodic repeat units of epithelial cell tonofilaments. *J. Cell. Biol.* **32**, 227-231.

Kim, K. H., Rheinwald, J. G., and Fuchs, E. (1983). Tissue specificity of epithelial keratins: Differential expression of mRNAs from two multigene families. *Mol. Cell. Biol.* **3**, 495-502.

Kim, K. H., Marchuk, D., and Fuchs, E. (1984). Expression of unusually large keratins during terminal differentiation: Balance of type I and type II keratins is not disrupted. *J. Cell Biol.* **99**, 1872-1877.

Kim, K. H., Schwartz, F., and Fuchs, E. (1984). Differences in keratin synthesis between normal epithelial cells and squamous cell carcinomas are mediated by vitamin A. *Proc. Natl. Acad. Sci. USA* **81**, 4280-4284.

Lazarides, E. (1982). Intermediate filaments; A chemically heterogeneous, developmentally regulated class of proteins. *Ann. Rev. Biochem.* **51**, 219-250.

Lane, E. B., Hogan, B. L. M., Kurkinen, M., and Garrels, J. I. (1983). Co-expression of vimentin and cytokeratins in parietal endoderm cells of early mouse embryo. *Nature* **303**, 701-704.

Lee, L. D., and Baden, H. P. (1976). Organisation of the polypeptide chains in mammalian keratin. *Nature* **264**, 377-379.

Lehnert, M. E., Jorcano, J. L., Zentgraf, H., Blessing, M., Franz, J. K., and Franke, W. W. (1984). Characterization of bovine keratin genes: similarities of exon patterns in genes coding for different keratins. *EMBO J.* **3**, 3279-3287.

Lewis, S. A., Balcarek, J. M., Krek, V., Shelanski, M., and Cowan, N. J. (1984). Sequence of a cDNA clone encoding mouse glial fibrillary acidic protein: structural conservation of intermediate filaments. *Proc. Natl. Acad. Sci. USA* **81**, 2743-2746.

Lonberg, N., and Gilbert, W. (1985). Intron/exon structure of the chicken pyruvate kinase gene. *Cell* **40**, 81-90.

Lu, Y.-J., and Johnson, P. (1983). The N-terminal domain of desmin is not involved in intermediate filament formation: evidence from thrombic digestion studies. *Int. J. Biol. Macromol.* **5**, 347-350.

Marchuk, D., McCrohon, S., and Fuchs, E. (1984). Remarkable conservation of structure among intermediate filament genes. *Cell* **39**, 491-498.

Marchuk, D., McCrohon, S., and Fuchs, E. (1985). Complete sequence of a gene encoding a human type I keratin: Sequences homologous to enhancer elements in the regulatory region of the gene. *Proc. Natl. Acad. Sci. USA* **82**, 1609-1613.

Markert, C. L., Shaklee, J. B., and Whitt, G. S. (1975). The evolution of a gene. *Science* **189**, 102-114.

McLachlan, A. D. (1978). Coiled coil formation and sequence regularities in the helical regions of α-keratin. *J. Mol. Biol.* **124**, 297-304.

McLachlan, A. D., and Stewart, M. (1982). Periodic charge distribution in the intermediate filament proteins desmin and vimentin. *J. Mol. Biol.* **162**, 693-698.

Milam, L., and Erickson, H. P. (1982). Visualization of a 21-nm axial periodicity in shadowed keratin filaments and neurofilaments. *J. Cell. Biol.* **94**, 592-596.

Moll, R., Franke, W. W., Schiller, D. L., Geiger, B., and Krepler, R. (1982). The catalog of human cytokeratins: Patterns of expression in normal epithelia, tumors and cultured cells. *Cell* **31**, 11-24.

Parry, D. A. D., Crewther, W. G., Fraser, R. D. B., and MacRae, T. P. (1977). Structure of α-keratin: Structural implication of the amino acid sequences of the type I and type II chain segments. *J. Mol. Biol.* **113**, 449-454.

Pauling, L., and Corey, R. B. (1953). Compound helical configurations of polypeptide chains: structure of proteins of the α-keratin type. *Nature* **171**, 59-61.

Quax, W., Egberts, W. V., Hendrisk, W., Quax-Jeuken, Y., and Bloemendal, H. (1983). The structure of the vimentin gene. *Cell* **35**, 215-223.

Quax, W., Heuvel, R.V.D., Egberts, W.V., Quax-Jeuken, Y., and Bloemendal H. (1984). Intermediate filament cDNAs from BHK-21 cells: demonstration of distinct genes for desmin and vimentin in all vertebrate classes. *Proc. Natl. Acad. Sci. USA* **81**, 5970-5974.

Quinlan, R. A., and Franke, W. W. (1982). Heteropolymer filaments of vimentin and desmin in vascular smooth muscle tissue and cultured baby hamster kidney cells demonstrated by chemical crosslinking. *Proc. Natl. Acad. Sci. USA* **79**, 3452-3456.

Quinlan, R. A., Cohlberg, J. A., Schiler, D. L., Hatzfeld, M., and Franke, W.W. (1984). Heterotypic tetramer (A2D2) complexes of non-epidermal keratins isolated from cytoskeletons of rat hepatocytes and hepatoma cells. *J. Mol. Biol.* **178**, 365-388.

Regier, J. C., Kafatos, F. C., and Hamodrakas, S.J. (1983). Silkmoth chorion multigene families constitute a superfamily: Comparison of C and B family sequences. *Proc. Natl. Acad. Sci. USA* **80**, 1043-1047.

Sharp, G., Osborn, M., and Weber, K. (1982). Occurrence of two different intermediate filament protein in the same filament *in situ* within a human glioma cell line. *Ex. Cell Res.* **141**, 385-395.

Skerrow, D., Matoltsy, G., and Matoltsy, M. (1973). Isolation and characterization of the helical regions of epidermal prekeratin. *J. Biol. Chem.* **248**, 4820-4826.

Steinert, P. M. (1978). Structure of the three-chain unit of the bovine epidermal keratin filaments. *J. Mol. Biol.* **123**, 49-70.

Steinert, P. M., Idler, W. W., and Zimmerman, S. B. (1976). Self-assembly of bovine epidermal keratin filaments *in vitro*. *J. Mol. Biol.* **108**, 547-467.

Steinert, P. M., Idler, W. W., and Goldman, R. D. (1980). Intermediate filaments of baby hamster kidney (BHK-21) cells and bovine epidermal keratinocytes have similar ultrastructures and subunit domain structures. *Proc. Natl. Acad. Sci. USA* **77**, 4534-4538.

Steinert, P., Idler, W., Aynardi-Whitman, M., Zackroff, R., and Goldman, R. D. (1982). Heterogeneity of intermediate filaments assembled *in vitro*. *Cold Spring Harbor Symp. Quant. Biol.* **46**, 465-474.

Steinert, P. M., Rice, R. H., Roop, D. R., Trus, B. L., and Steven, A. C. (1983). Complete amino acid sequence of a mouse epidermal keratin subunit and implications for the structure of intermediate filaments. *Nature* **302**, 794-800.

Steinert, P. M., Parry, D. A. D., Racoosin, E. L., Idler, W. W., Steven, A. C., Trus, B. L., and Roop, D. R. (1984). The complete cDNA and deduced amino acid sequence of a type II mouse epidermal keratin of 60,000 Da: analysis of sequence differences between type I and type II keratins. *Proc. Natl. Acad. Sci. USA* **81**, 5709-5713.

Sun, T.-T., Shih, C., and Green, H. (1979). Keratin cytoskeletons in epithelial cells of internal organs. *Proc. Natl. Acad. Sci. USA* **76**, 2813-2817.

Wahli, W., Dawid, I. B., Ryfell, G. U., and Weber, R. (1981). Vitellogenesis and vitellogenin gene family. *Science* **212**, 298-304.

Walter, M. F., and Biessmann, H. (1984). A monoclonal antibody that detects vimentin-related proteins in invertebrates. *Mol. Cell. Biochem.* **60**, 99-108.

Wang, E., Cairncross, J. G., and Liem, R. K. H. (1984). Identification of glial filament protein and vimentin in the same intermediate filament system in human glioma cells. *Proc. Natl. Acad. Sci. USA* **81**, 2102-2106.

Weber, K., Osborn, M., and Franke, W. W. (1980). Antibodies against merokeratin from sheep wool decorate cytokeratin filaments in non-keratinizing epithelial cells. *Eur. J. Cell. biol.* **23**, 110-114.

Woods, E. F., and Gruen, L. C. (1983). Sturctural studies on the microfibrillar proteins of wool: characterization of the α-helix particle produced by chymotryptic digestion. *Aust. J. Biol. Sci.* **34**, 515-526.

Zehner, Z. E., and Paterson, B. M. (1983). Characterization of the chicken vimentin gene: single copy gene producing multiple mRNAs. *Proc. Natl. Acad. Sci. USA* **80**, 911-915.

Note added in proof:

A recent paper has reported partial sequence homologies between some IF proteins and oncogene proteins (Crabbe, 1985). In addition, the recently determined sequence of the Scrapie protein reveals Gly-Gly-X type repeats that are also found in cytoskeletal keratins (Oesch *et al.*, 1985).

Crabbe, M. J. C. (1985). Partial sequence homologies between cytoskeletal proteins, c-myc, Rous sarcoma virus and adenovirus proteins, transducin, and β- and γ- crystallins. *Biosci. Rep.* **5**, 167-174.

Oesch, B., Westaway, D., Walchli, M., McKinley, M. P., Kent, S. B. H., Aebersold, R., Barry, R. A., Tempst, P., Teplow, D. B., Hood, L. E., Prusiner, S. B., and Weissmann, C. (1985). A cellular gene encodes scrapie PrP 27-30 protein. *Cell* **40**, 735-746.

After the completion of this manuscript additional sequences of intermediate filament protein genes have been reported (see references below). The characteristics of these sequences are consistent with the conclusions noted in this review.

Geisler, N. Plessmann, U., and Weber, K. (1985). The complete amino acid sequence of the major mammalian neurofilament protein (NF-L). *FEBS Lett.* **182**, 475-478.

Johnson, D. L., Idler, W. W., Zhou, X.-M., Roop, D. R., and Steinert, P. M. (1985). Structure of a gene for the human epidermal 67-kDa keratin. *Proc. Natl. Acad. Sci. USA* **82**, 1896-1900.

Jorcano, J. L., Franz, J. K., and Franke, W. W. (1984). Amino acid sequence diversity between bovine epidermal cytokeratin polypeptides of the basic (type II) subfamily as determined from cDNA clones. *Differentiation* **28**, 155-163.

Krieg, T. M., Schafer, M. P., Cheng, C. K., Filpula, D., Flaherty, P., Steinert, P. M., and Roop, D. R. (1985). Organization of a type I keratin gene. *J. Biol. Chem.* **260**, 5867-5870.

Steinert, P. M., Parry, D.A.D., Idler, W. W., Johnson, D. L., Steven, A. C., and Roop, D. R. (1985). Amino acid sequences of mouse and human epidermal type II keratins of Mr 67,000 provide a systematic basis for the structural and functional diversity of the end domains of keratin intermediate filament subunits. *J. Biol. Chem.* **260**, 7142-7149.

Tyner, A. L., Eichman, M. J., and Fuchs, E. (1985). The sequence of a type II keratin gene expressed in human skin: Conservation of structure among all intermediate filament genes. *Proc. Natl. Acad. Sci. USA* **82**, 4683-4687.