Cell, Vol. 31, 243-252, November 1982, Copyright © 1982 by MIT

The cDNA Sequence of a Human Epidermal Keratin: Divergence of Sequence but Conservation of Structure among Intermediate Filament Proteins

Israel Hanukoglu and Elaine Fuchs
Department of Biochemistry
The University of Chicago
Chicago, Illinois 60637

Summary

We have determined the DNA sequence of a cloned cDNA that is complementary to the mRNA for the 50 kilodalton (kd) human epidermal keratin. This provides the first amino acid sequence for a cytoskeletal keratin. Comparison of this sequence with those of other keratins reveals an evolutionary relationship between the cytoskeletal and the microfibrillar keratins, but shows no homology to matrix or feather keratins. The 50 kd keratin shares 28%–30% homology with partial sequences of other intermediate filament proteins, which suggests that keratins may be the most distantly related members of this class of fibrous proteins. Our computer analyses predict that the 50 kd keratin contains two long α-helical domains separated by a cluster of helix-inhibitory residues in the middle of the protein. These findings indicate that despite major sequence divergence among intermediate filament proteins, they retain sequences compatible with secondary structural features that appear to be common to all of them.

Introduction

Intermediate filaments (7–10 nm diameter) are a major component of the cytoskeleton of most mammalian cells. The polypeptide subunits that constitute these filaments have been grouped into five categories: keratins, desmin, vimentin, neurofilament proteins and glial filament proteins (Lazarides, 1980). Within a particular cell, the ratios and the spectrum of these subunits can vary during the course of differentiation (Fuchs and Green, 1980; Gard and Lazarides, 1980; Schmid et al., 1982).

Recently, several studies have suggested biochemical, immunological and structural similarities among all intermediate filament proteins (Steinert et al., 1980; Pruss et al., 1981). Moreover, amino acid sequence analyses of short fragments of four porcine intermediate filament proteins have revealed significant homologies between desmin, vimentin and a neurofilament protein, and have provided preliminary evidence that indicates some homology between a glial filament protein and the other three proteins (Geisler et al., 1982). Therefore, although amino acid sequence data is not yet available for a cytoskeletal keratin, some sequence homologies between keratins and other intermediate filaments could be expected.

Among the five categories of intermediate filament subunits, keratins represent the largest and most diverse class. They constitute a group of 10–20 closely related proteins (40–70 kd) that are found in most epithelial cells (Baden et al., 1973; Steinert and Idler, 1975; Brysk et al., 1977; Culbertson and Freedberg, 1977; Franke et al., 1978, 1979; Fuchs and Green, 1978; Sun et al., 1979). In addition to these cytoskeletal keratins, many proteins derived from epidermal appendages such as hair, wool and feather, have been termed keratins mainly for historical reasons (Fraser et al., 1972). The mammalian epidermal appendages consist of microfibrils embedded in a rigid matrix of other proteins. The structure of the microfibrils appears similar to cytoskeletal intermediate filaments. The proteins that are the presumed building blocks of the microfibrils are referred to as "low-sulfur keratins" and have a range of 40–56 kd (Fraser et al., 1972; Jones, 1975). The matrix proteins, which are also called keratins, are mostly small proteins (6–20 kd). While the primary sequences of some of the matrix keratins have already been determined (for example, Swart and Haylett, 1973), only partial amino acid sequences of microfibrillar keratins have been obtained (Gough et al., 1978), and no sequence information is available for the cytoskeletal keratins. Thus the relationship of the cytoskeletal keratins to these other keratins has remained largely a matter of conjecture.

The significance of the five groups of similar intermediate filament proteins and of the multiplicity of proteins within a particular group, such as the keratins, has not yet been elucidated. As some reconstitution studies indicate, some of the heterogeneity of polypeptide subunits appears to be essential for filament assembly (Lee and Baden, 1976; Steinert et al., 1976). It seems evident, however, that further understanding of the complexities of intermediate filament subunits requires a more detailed analysis of their structural organization. Although structural models of intermediate filaments and epidermal appendage microfibrils have been advanced (for example, Crick, 1953; Pauling and Corey, 1953; Skerrow et al., 1973; Fraser et al., 1976; Steinert, 1978), these need to be further tested and refined with quantitative models based on the primary structures of the individual subunits.

Recently we cloned the cDNAs of epidermal keratins. Analysis of these clones revealed at least two major types of keratin mRNAs that are coordinately conserved throughout vertebrate evolution, and which may be essential for keratin filament assembly (Fuchs et al., 1981). In view of the need to understand the molecular architecture of the filaments, we have sequenced two clones that belong to one of these two mRNA classes. We present the predicted primary and secondary structures of the protein coded by these sequences, and compare its structural properties with those of other keratins and intermediate filament proteins. This represents the first presentation of the nearly complete sequence of an intermediate filament protein.

Results

cDNAs Sequenced and DNA Sequencing Strategy

The cDNAs of keratin mRNAs present in cultured human epidermal cells have been cloned with pBR322 as vector and Escherichia coli X 1776 as host (Fuchs et al., 1981). The hybrid plasmids were constructed by the insertion of double-stranded cDNA into the Pst I site of pBR322. Within a collection of cloned hybrid pBR322–cDNA plasmids, two classes of keratin cDNAs were identified by positive hybrid-selection and translation. One class hybridized with mRNAs that code for 56 kd and 58 kd keratins, and the other class hybridized with mRNAs that code for 46 kd and 50 kd keratins (Fuchs et al., 1981). We selected two different cloned cDNAs, KB-2 and KB-64, that belong to the latter class. KB-2 was selected because it represented the cDNA insert nearest in size to the complete mRNA from which it was derived (Fuchs et al., 1981), and KB-64 was selected for comparative reasons outlined below. More recent experiments with smaller restriction fragments of KB-2 cDNA for positive hybrid-selection indicate that the insert of this clone codes for the 50 kd keratin rather than the 46 kd keratin (K. H. Kim and E. Fuchs, manuscript in preparation).

The strategies used in determining the DNA sequence of cDNA inserts KB-2 and KB-64 are shown in Figure 1. The only uncertainties in the sequencing involved Eco RII sites in which the middle cytosine was methylated by an Eco RII methylase (Razin et al., 1980), resulting in a blank line in the nucleotide ladder. However, as indicated in Figure 1, we sequenced both strands of the cDNA insert, and could thus confirm the identification of the methylated cytosines. Our initial restriction maps indicated that KB-64 had the same sites as KB-2, with the exception that the ends of KB-64 were approximately 40 and 200 nucleotides shorter than those of KB-2. We determined the sequences of both, as a control for artifactual inverted sequences or single-base-pair copying errors that can be generated during the preparation of cDNAs for cloning (Weaver et al., 1981). Thus the sequencing of both clones increased our confidence in the fidelity of the cDNA sequence as a copy of the mRNA.

cDNA Sequences

The DNA sequence of the KB-2 keratin cDNA insert is shown in Figure 2. The size of the insert is 1421 nucleotides. The DNA sequence of the shorter insert, KB-64, was found to be identical to the corresponding sequenced region of KB-2, with the exception of the one nucleotide indicated in the legend. The cDNA insert KB-2 included a copy of a portion of the poly(A) tail of the 3′ end of the mRNA. In addition to the 9 nucleotide poly(A) segment, the sequences 5′-AAUACA-3′ and 5′-AAUCAA-3′ were found 21 and 25 nucleotides, respectively, from the first A of the poly(A). A similar sequence, 5′-AAUAAA-3′, has been found approximately 20 nucleotides from the poly(A) tail initiation site in almost all polyadenylated RNA sequences, and it may serve as a signal for polyadenylation (Fitzgerald and Shenk, 1981). Additional variants of this sequence have been observed in other eucaryotic cDNA sequences, which indicates some flexibility in sequence recognition at this site (Valenzuela et al., 1981).

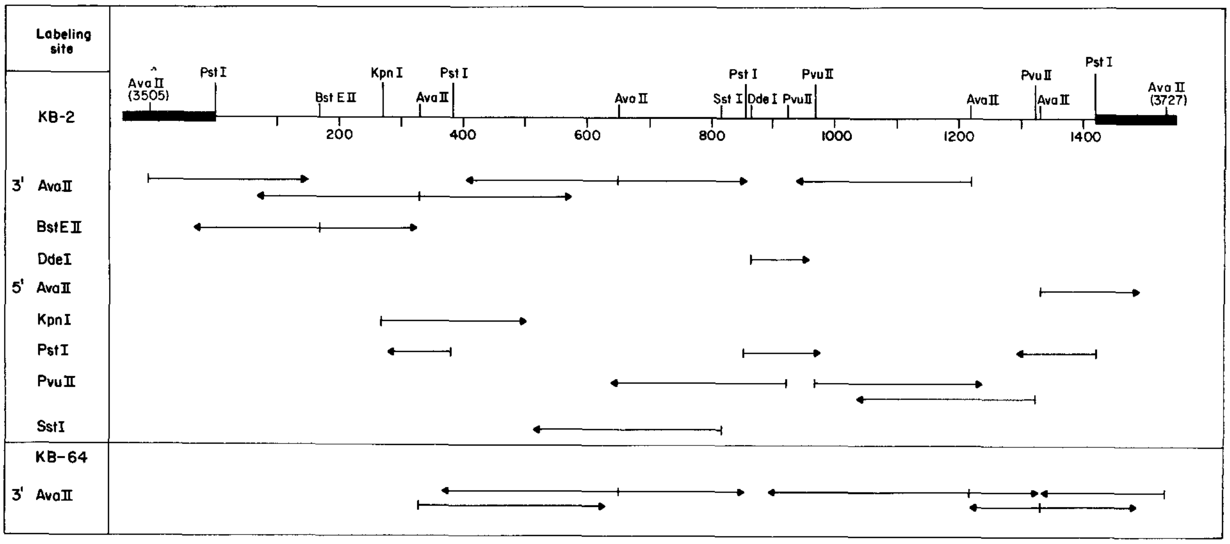

Figure 1. DNA Sequencing Strategies for the Human Epidermal 50 kd Keratin cDNA Inserts KB-2 and KB-64

The KB-2 cDNA insert (thin line) flanked by pBR322 sequences (heavy lines) is shown at the top. The nucleotide numbers within the KB-2 insert are in the 5′-to-3′ direction of the mRNA strand, and the positions of all recognition sites for each enzyme, except Dde I, are indicated. The ^{32}P labeling site for each series of restriction fragments is shown at the left. Arrows: direction and extent of DNA sequence determination.

```
                                                                                                                (G)^9AC
                                     10                                      20                                      30
Gly Leu Gly Gly Gly Tyr Gly Gly Gly Phe Ser Ser Ser Ser Ser Ser Phe Gly Ser Gly Phe Gly Gly Gly Tyr Gly Gly Gly Leu Gly
GGG CTG GGG GGC GGC TAT GGC GGT GGC TTC AGC AGC AGC AGT AGC AGC TTT GGT AGT GGC TTT GGG GGA GGA TAT GGT GGT GGC CTT GGT
                                     30                                      60                                      90

                                     40                                      50                                      60
Thr Gly Leu Gly Gly Gly Phe Gly Gly Gly Phe Ala Gly Gly Asp Gly Leu Leu Val Gly Ser Glu Lys Val Thr Met Gln Asn Leu Asn
ACT GGC TTG GGT GGT GGC TTT GGT GGT GGC TTT GCT GGT GGT GAT GGG CTT CTG GTG GGC AGT GAG AAG GTG ACC ATG CAG AAC CTC AAT
                                    120                                     150                                     180

                                     70                                      80                                      90
Asp Arg Leu Ala Ser Tyr Leu Asp Lys Val Arg Ala Leu Glu Glu Ala Asn Ala Asp Leu Glu Val Lys Ile Arg Asp Trp Tyr Gln Arg
GAC CGC CTG GCC TCC TAC CTG GAC AAG GTG CGT GCT CTG GAG GAG GCC AAC GCC GAC CTG GAA GTG AAG ATC CGT GAC TGG TAC CAG AGG
                                    210                                     240                                     270

                                    100                                     110                                     120
Gln Arg Pro Ala Glu Ile Lys Asp Tyr Ser Pro Tyr Phe Lys Thr Ile Glu Asp Leu Arg Asn Lys Ile Leu Thr Ala Thr Val Asp Asn
CAG CGG CCT GCT GAG ATC AAA GAC TAC AGT CCC TAC TTC AAG ACC ATT GAG GAC CTG AGG AAC AAG ATT CTC ACA GCC ACA GTG GAC AAT
                                    300                                     330                                     360

                                    130                                     140                                     150
Ala Asn Val Leu Leu Gln Ile Asp Asn Ala Arg Leu Ala Ala Asp Asp Phe Arg Thr Lys Tyr Glu Thr Glu Leu Asn Leu Arg Met Ser
GCC AAT GTC CTT CTG CAG ATT GAC AAT GCC CGT CTG GCC GCG GAT GAC TTC CGC ACC AAG TAT GAG ACA GAG TTG AAC CTG CGC ATG AGT
                                    390                                     420                                     450

                                    160                                     170                                     180
Val Glu Ala Asp Ile Asn Gly Leu Arg Arg Val Leu Asp Glu Leu Thr Leu Ala Arg Ala Asp Leu Glu Met Gln Ile Glu Ser Leu Lys
GTG GAA GCC GAC ATC AAT GGC CTG CGC AGG GTG CTG GAC GAA CTG ACC CTG GCC AGA GCT GAC CTG GAG ATG CAG ATT GAG AGC CTG AAG
                                    480                                     510                                     540

                                    190                                     200                                     210
Glu Glu Leu Ala Tyr Leu Lys Lys Asn His Glu Glu Glu Met Asn Ala Leu Arg Gly Gln Val Gly Gly Asp Val Asn Val Glu Met Asp
GAG GAG CTG GCC TAC CTG AAG AAG AAC CAC GAG GAG GAG ATG AAT GCC CTG AGA GGC CAG GTG GGT GGA GAT GTC AAT GTG GAG ATG GAC
                                    570                                     600                                     630

                                    220                                     230                                     240
Ala Ala Pro Gly Val Asp Leu Ser Arg Ile Leu Asn Glu Met Arg Asp Gln Tyr Glu Lys Met Ala Glu Lys Asn Arg Lys Asp Ala Glu
GCT GCA CCT GGC GTG GAC CTG AGC CGC ATT CTG AAC GAG ATG CGT GAC CAG TAT GAG AAG ATG GCA GAG AAG AAC CGC AAG GAT GCC GAG
                                    660                                     690                                     720

                                    250                                     260                                     270
Glu Trp Phe Phe Thr Lys Thr Glu Glu Leu Asn Arg Glu Val Ala Thr Asn Ser Glu Leu Val Gln Ser Gly Lys Ser Glu Ile Ser Glu
GAA TGG TTC TTC ACC AAG ACA GAG GAG CTG AAC CGC GAG GTG GCC ACC AAC AGC GAG CTG GTG CAG AGC GGC AAG AGC GAG ATC TCG GAG
                                    750                                     780                                     810

                                    280                                     290                                     300
Leu Arg Arg Thr Met Gln Asn Leu Glu Ile Glu Leu Gln Ser Gln Leu Ser Met Lys Ala Ser Leu Glu Asn Ser Leu Glu Glu Thr Lys
CTC CGG CGC ACC ATG CAG AAC CTG GAG ATT GAG CTG CAG TCC CAG CTC AGC ATG AAA GCA TCC CTG GAG AAC AGC CTG GAG GAG ACC AAA
                                    840                                     870                                     900

                                    310                                     320                                     330
Gly Arg Tyr Cys Met Gln Leu Ala Gln Ile Gln Glu Met Ile Gly Ser Val Glu Glu Gln Leu Ala Gln Leu Arg Cys Glu Met Glu Gln
GGT CGC TAC TGC ATG CAG CTG GCC CAG ATC CAG GAG ATG ATT GGC AGC GTG GAG GAG CAG CTG GCC CAG CTC CGC TGC GAG ATG GAG CAG
                                    930                                     960                                     990

                                    340                                     350                                     360
Gln Asn Gln Glu Tyr Lys Ile Leu Leu Asp Val Lys Thr Arg Leu Glu Gln Glu Ile Ala Thr Tyr Arg Arg Leu Leu Glu Gly Glu Asp
CAG AAC CAG GAG TAC AAG ATC CTG CTG GAC GTG AAG ACG CGG CTG GAG CAG GAG ATC GCC ACC TAC CGC CGC CTG CTG GAG GGC GAG GAC
                                   1020                                    1050                                    1080

                                    370                                     380                                     390
Ala His Leu Ser Ser Ser Gln Phe Ser Ser Gly Ser Gln Ser Ser Arg Asp Val Thr Ser Ser Ser Arg Gln Ile Arg Thr Lys Val Met
GCC CAC CTC TCC TCC TCC CAG TTC TCC TCT GGA TCG CAG TCA TCC AGA GAT GTG ACC TCC TCC AGC CGC CAA ATC CGC ACC AAG GTC ATG
                                   1110                                    1140                                    1170

                                    400                                     410
Asp Val His Asp Gly Lys Val Val Ser Thr His Glu Gln Val Leu Arg Thr Lys Asn  .
GAT GTG CAC GAT GGC AAG GTG GTG TCC ACC CAC GAG CAG GTC CTT CGC ACC AAG AAC TGA GGCTGCCCAGCCCCGCTCAGGCCTAGGAGGCCCCCCGTG
                                   1200                                    1230                           1260

TGGACACAGATCCCACTGGAAGATCCCCTCTCCTGCCCAAGCACTTCACAGCTGGACCCTGCTTCACCCTCACCCCCTCCTGGCAATCAATACAGCTTCATTATCTGAGTTGCATAAAA
                 1290                          1320                          1350                          1380

AAAAA(C)^17
```

Figure 2. DNA Sequence of the KB-2 cDNA Insert and Predicted Amino Acid Sequence of the Human Epidermal 50 kd Keratin

The sequence is shown in the 5′-to-3′ direction of the mRNA strand. The numbers above the amino acids mark the positions of the amino acids, and those below mark the positions of the nucleotides (N-terminus of the predicted amino acid sequence = 1). The cluster of Gs at the 5′ end and the cluster of Cs at the 3′ end represent the enzymatically tailed regions of the plasmid and the double-stranded cDNA used for cloning (Fuchs et al., 1981). The stop codon of the reading frame is marked with a dot. Underlined nucleotides: putative polyadenylation signal sequence. The nucleotide at position 559 is a G instead of an A in the KB-64 cDNA. This converts the codon for Lys to that for Glu.

With the reading frame shown in Figure 2, it appears that the mRNA contains a 3′ end untranslated region that begins at the stop codon TGA, which occurs 155 nucleotides from the poly(A) segment. There is only one other stop codon in the same reading frame, which starts 10 nucleotides before the poly(A) segment. The 5′ end of the cDNA does not contain a complete copy of the 5′ end of the mRNA, since at the beginning of the translated sequence there is no translation initiation codon for methionine and no untranslated 5′ end region (Figure 2).

We screened for the presence of repetitive or palindromic sequences with the algorithms available on the SEQ-DNA sequence analysis system (Brutlag et al., 1982). This search revealed the presence of statistically significant homologies in the first approximately 150 nucleotides of the KB-2 sequence (for example, segment 50–88 with segment 86–124; Figure 2). Downstream from this region, eight segments were found to show significant homologies: 163–176 with 820–833, 225–251 with 501–527, 527–557 with 939–971 and 915–931 with 957–973. In addition to these homologies, there were six statistically significant symmetric regions (palindromic repeats); these are 83–64 with 100–119, 107–87 with 112–132 and 1048–1032 with 1065–1081.

Predicted Amino Acid Sequence of a Human Epidermal Keratin

The predicted amino acid sequence of the human epidermal keratin encoded by the cDNA insert KB-2 is shown in Figure 2. Since the amino acid sequence of an epidermal keratin from any species has never been determined, we first assigned as the sense strand the one shown, according to the presence of its poly(A) stretch and the polyadenylation signal sequence. The other two reading frames of the sense strand (not shown) were both found to contain 17 stop codons scattered randomly throughout the sequence, indicating that they do not code for a functional protein.

As previously noted, a restriction fragment of the cloned KB-2 insert specifically hybridized an mRNA, which translated in vitro into a 50 kd protein. One-dimensional peptide mapping was used to verify this further as the 50 kd keratin (Fuchs et al., 1981). The molecular weight of the protein segment coded by the insert of KB-2 is 46 kd. Thus if the electrophoretic estimation of the molecular weight of the 50 kd keratin is accurate, the KB-2 insert represents 92% of the complete sequence of this keratin. To confirm the identity of the predicted translation product, we isolated the 50 kd keratin from cultured human epidermal cells and purified it, and determined its amino acid composition for comparison with the predicted values. The amino acid composition of the predicted protein segment is very similar to that determined for the 50 kd keratin (Table 1, columns 6–7), particularly with respect to the unusually high Glu and Gln composition of the protein. This provides further evidence that the insert of KB-2 codes for this keratin. The strand complementary to the sequence shown in Figure 2 contains one open reading frame over 1000 nucleotides long; however, the amino acid composition of this reading frame differs markedly from that of the 50 kd keratin. Furthermore, as previously indicated, this other strand does not contain a poly(A) segment or a putative polyadenylation signal. Collectively, these results clearly establish that the correct reading frame for the KB-2 insert is the one shown in Figure 2, and that this insert represents an accurate copy of the mRNA that codes for the 50 kd epidermal keratin.

Non-Random Amino Acid Codon Frequencies for the Keratin cDNA Sequence

Within the reading frame shown in Figure 2, the codon frequencies of many amino acids are not random (Table 2). Most notably, there is a distinct predominance of codons ending with G or C (Table 2). For example, while 18 codons for valine end in G, only four end in C and none end in A or T; similarly for glutamate and glutamine, the codons ending in G occur 10-fold and 23-fold more than the respective alternative codons (Table 2). Although a similar bias has been noted in some human cDNAs (for references, see Fiddes and Goodman, 1980), the magnitude of the bias is significantly greater for the keratin cDNA sequence. There is evidence that in animal genes, codons ending in G or C are preferred (Wain-Hobson et al., 1981); however, in at least one human mRNA there is no preference for G or C in the third codon position (Hendy et al., 1981).

Homologies between the 50 kd Keratin Sequence and the Sequences of Other Keratins

The amino acid composition and the molecular weight of the 50 kd epidermal keratin are similar to those of the microfibrillar keratins of hair and wool, but markedly different from those of the matrix keratins (Table

Table 1. The Amino Acid Composition of the 50 kd Human Epidermal Keratin as Compared to Other Types of Keratins

		Matrix (Wool)		Microfibrillar		Epidermal 50 kd	
Amino Acid	Emu Feather[a]	High-Sulfur	High-Tyrosine	Rat	Wool[b]	Amino Acid Analysis	cDNA[c]
Ala	3.9	2.3	3.3	7.5	6.4	7.7	6.4
Arg	3.9	10.3	3.3	6.7	7.3	6.1	6.8
Asn	4.9	0.0	0.0				4.9
Asp	1.0	1.5	1.6				5.9
Asn + Asp	5.9	1.5	1.6	8.6	8.1	9.6	10.8
Cys	7.8	24.5	6.6	5.3	6.8	n.d.	0.5
Gln	3.9	5.0	0.0				5.9
Glu	2.0	1.5	0.0				11.0
Gln + Glu	5.9	6.5	0.0	13.4	14.1	15.6	16.9
Gly	10.8	4.6	23.0	8.3	8.8	11.8	9.3
His	0.0	0.0	0.0	1.8	0.7	1.0	1.0
Ile	3.9	1.5	0.0	3.9	3.7	3.8	3.7
Leu	8.8	2.3	3.3	8.8	10.3	10.6	11.2
Lys	0.0	0.0	0.0	5.4	4.1	4.8	5.4
Met	0.0	0.0	0.0	1.0	0.6	2.4	3.2
Phe	2.9	2.3	9.8	3.0	3.0	3.2	2.4
Pro	11.8	14.2	6.6	4.8	4.2	1.3	0.7
Ser	18.6	10.0	13.1	7.9	7.3	9.7	8.3
Thr	4.9	11.1	3.3	5.6	4.4	4.8	4.6
Trp	0.0	0.8	3.3	n.d.	n.d.	n.d.	0.5
Tyr	2.0	1.1	18.0	2.5	4.3	3.0	2.9
Val	8.8	6.9	4.9	5.7	5.9	4.9	5.4
Residues sequenced	102	131	61				>410
Weight (daltons)	10,293	14,164	6,660	45,000–50,000	45,000–50,000	50,000	50,000

Values are presented as percentages.

[a] Determined as described in the Experimental Procedures.

[b] From O'Donnell, 1973; Swart and Haylett, 1973; and Jones, 1975.

[c] Calculated from the predicted protein sequence in Figure 2.

Table 2. The Frequency of Amino Acid Codons in the Coding Region of the KB-2 cDNA Insert

TTT	Phe	4	(1.0)	TCT	Ser	1	(0.2)	TAT	Tyr	4	(1.0)	TGT	Cys	0	(0.0)
TTC	Phe	6	(1.5)	TCC	Ser	11	(2.7)	TAC	Tyr	8	(2.0)	TGC	Cys	2	(0.5)
TTA	Leu	0	(0.0)	TCA	Ser	1	(0.2)	TAA	Ter.	0	(0.0)	TGA	Ter.	0	(0.0)
TTG	Leu	2	(0.5)	TCG	Ser	2	(0.5)	TAG	Ter.	0	(0.0)	TGG	Trp	2	(0.5)
CTT	Leu	4	(1.0)	CCT	Pro	2	(0.5)	CAT	His	0	(0.0)	CGT	Arg	4	(1.0)
CTC	Leu	6	(1.5)	CCC	Pro	1	(0.2)	CAC	His	4	(1.0)	CGC	Arg	15	(3.7)
CTA	Leu	0	(0.0)	CCA	Pro	0	(0.0)	CAA	Gln	1	(0.2)	CGA	Arg	0	(0.0)
CTG	Leu	34	(8.3)	CCG	Pro	0	(0.0)	CAG	Gln	23	(5.6)	CGG	Arg	3	(0.7)
ATT	Ile	7	(1.7)	ACT	Thr	1	(0.2)	AAT	Asn	7	(1.7)	AGT	Ser	5	(1.2)
ATC	Ile	8	(2.0)	ACC	Thr	13	(3.2)	AAC	Asn	13	(3.2)	AGC	Ser	14	(3.4)
ATA	Ile	0	(0.0)	ACA	Thr	4	(1.0)	AAA	Lys	3	(0.7)	AGA	Arg	3	(0.7)
ATG	Met	13	(3.2)	ACG	Thr	1	(0.2)	AAG	Lys	19	(4.6)	AGG	Arg	3	(0.7)
GTT	Val	0	(0.0)	GCT	Ala	5	(1.2)	GAT	Asp	7	(1.7)	GGT	Gly	13	(3.2)
GTC	Val	4	(1.0)	GCC	Ala	17	(4.2)	GAC	Asp	17	(4.2)	GGC	Gly	17	(4.2)
GTA	Val	0	(0.0)	GCA	Ala	3	(0.7)	GAA	Glu	4	(1.0)	GGA	Gly	4	(1.0)
GTG	Val	18	(4.4)	GCG	Ala	1	(0.2)	GAG	Glu	41	(10.0)	GGG	Gly	4	(1.0)

Numbers in parentheses: percentage of each codon.

1). On the basis of their similarities, some homology might be expected between the 50 kd epidermal keratins and the microfibrillar keratin. The longest fragments of microfibrillar keratins for which the amino acid sequences are known are about 100 residues long and were derived from wool (Crewther et al., 1978; Gough et al., 1978). Our results show that a segment of the 50 kd keratin is homologous to these two protein fragments (Figure 3), with 59% and 27% homology for the "type-I" and "type-II" fragments, respectively.

The complete sequences for several of the "high sulfur" matrix keratins of wool have been determined (for example, Swart and Haylett, 1973). A comparison of the sequence of the 50 kd epidermal keratin with these matrix keratin sequences revealed no significant homology between the two types of proteins. In contrast, up to 20% homology was observed between a small glycine- and tyrosine-rich matrix keratin sequence (see Table 1; Dopheide, 1973) and the glycine-rich amino-terminal region of the 50 kd when these sequences were aligned in several different frames. However, since the tandemly repeated pattern of glycines in the amino-terminal region of the 50 kd keratin is markedly different from the distribution of glycines in this matrix protein, the relatively high degree of homology could be ascribed to the glycine richness of the two proteins rather than to a true homology.

Homologies between the 50 kd Keratin Sequence and the Sequences of Other Intermediate Filament Proteins

The sequences of short fragments of three porcine intermediate filament proteins, desmin, vimentin and a neurofilament protein, have been determined (Geisler and Weber, 1981; Geisler et al., 1982). These studies revealed 42% homology of the neurofilament protein with both desmin and vimentin, and about 70% homology between desmin and vimentin. Although segments of the 50 kd keratin sequence share significant homology with these three proteins (Figure 4), only 28%, 29% and 30% homology with desmin, vimentin and the neurofilament protein was observed, given the alignments of the sequences shown in Figure 4.

Predicted Secondary Structure of the 50 kd Keratin

We used two probabilistic methods for the prediction of the secondary structure of the 50 kd keratin based on its primary structure. The methods of both Chou and Fasman (1978) and Robson and coworkers (B. Robson, E. Suzuki, R. H. Pain and J. Garnier, University of Manchester, Manchester, England) (see Experimental Procedures) yield similar secondary structures for this protein (Figure 5). The most notable common feature is that, excluding the amino- and carboxy-terminal regions, large segments of the pro-

```
50K-WK1:             |  || ||| |  ||| |   |      |||| || |||||||| |  ||  || ||| ||| ||||||    ||   |||||||  ||||||||   |
50K-WK2:   |     |          | ||     |  |                     ||| |  ||   |     |    |   |||   | |  |   |    |||   |
50K    : 89QRQRPAEIKDYSPYFKTIEDLRNKILTATVDNANVLLQIDNARLAADDFRTKYETELNLRMSVEADINGLRRVLDELTLARADLEMQIESLKEELAYLKKNHEEEMNAL197
WK1    :       LC  PNYQSYFRTIEELQQKILCAKSENSRLVIEIDNAKLASDDFRTKYESERSLRQLVESDINSLRRILDELTLCKSNLEAEVESLKEELLCLKKNHEEEADSL
WK2    :   QNRQCCESNLEPLFSGYIETLRREAECAEADSGRLSSELNSLQEVLEGYKKKYEEEIALRATAENEFVALKKDVDCAYLRKSDLEANVEALIQETDFLRRLYEEEIRVL
WK1-WK2:        |           || |     ||     ||  |                ||| |  ||   |     |    |   | || ||| || |  |   |    |||   |
```

Figure 3. Amino Acid Sequence Homologies between the 50 kd Human Epidermal Keratin and Type-I (WK1) and Type-II (WK2) Segments of Microfibrillar Keratins from Wool

The wool keratin sequences are from Gough et al. (1978). Vertical bars: positions of homology between the two sequences indicated at the left side of each line. The percentage of homologous positions for the three comparisons are 50K–WK1: 59%. 50K–WK2: 27%. WK1–WK2: 30%. The two blank spaces in the WK1 sequence were introduced to maximize the homology between WK1 and WK2. The numbers to the right and left of the 50K sequence represent the positions of the first and last amino acids of the segment in Figure 2.

```
K-D:    |   |   |      |       | |  | |   |    |         ||     |   |        |  |   ||    |  ||     ||  || ||  |  |||||| |||||
K-V:    |   |   |      |       | |  | ||  | |  |         |||    |         |  |  ||           ||     ||  || ||  |  |||||| |||||
K  : 242WFFTKTEELNREVATNSELVQSGKSEISELRRTMQNLEIELQSQLSMKASLENSLEETKGRYCMQLAQIQEMIGSVEEQLAQLRCEMEQQNQEYKILLDVKTRLEQEIATYRRLLEGED360
D  :    WYKSKVSDLTQAANKNNDALRQAKQEMMEYRHEIQSYTCEIDALKGTNDSLMRQMRELEDRFASEASGYQDNIARLEEEIRHLKDEMARHLREYQDLLNVKMALDVEIATYRKLLEGEE
V  :    WYKSKFADLSEAANRNNDALRQAKQESNEYRRQVQSLTCEVDALKGTNESLERQMREMEENFAVEAANYQDTIGRLQDEIQNMKEEMARHLREYQDLLNVKMALDIEIATYRKLLEGEE
D-V:    |||||  ||  ||| |||||||||||  |||   || ||| ||||||| || ||||| |  || ||  ||| | ||  ||   | |||||||||||||||||||| |||||||||||||

K-D:           | ||                      |         || |||   |
K-V:       |     |||         |         ||          || |     |
K  : 361AHLSSS Q FSSGSQSSRDVTSSSRQI    RTKVMDVH   DGKVVSTHEQVLRTKN409
D  :    SRINLPIQTFSALNFRET SPEQRGSEVHTKKTVMIKTIETRDGEVVSEATQQQHEVL
V  :    SRISLPLPNFSSLNLRETNLESLPLVDTHSKRTLLIKTVETRDGQVINETSQHHNDLE
D-V:    ||| ||   || || |||          | | |  ||| ||||| |  |  |
```

Figure 4. Amino Acid Sequence Homologies between the 50 kd Human Epidermal Keratin, Porcine Desmin (D) and Vimentin (V)

The porcine desmin and vimentin sequences are from Geisler and Weber (1981) and Geisler et al. (1982). The percentage of homologous positions for the three comparisons are 50K-D: 28%. 50K-V: 29%. D-V: 66%. Other details are the same as in Figure 3. The homology between the 50K keratin and a porcine neurofilament protein fragment (Geisler et al., 1982) is 30%. The sequence of this fragment is not shown, since it represents less than 10% of the complete sequence of the protein.

tein are predicted to be in α-helical conformation. The magnitude of the propensity for this conformation is more clearly illustrated by the conformational profile of the amino acid sequence, which indicates that the probability for α-helical conformation is dominant over that for β-sheet conformation throughout most of the molecule (Figure 5).

Only the glycine-rich amino-terminal region of the protein contains long sequences that are not likely to form stable helical structures. However, it is unlikely that the helical conformation of the protein is propagated throughout the whole length of the molecule, since there is a cluster of three glycines close to a proline and a glycine in the middle of the protein (Figures 2 and 5; residues 199–214). A similar cluster of helix-inhibitory residues is found approximately 100 residues from the amino-terminal end of the molecule (Figures 2 and 5).

Another feature of the secondary structure analysis is that very few β-turns are predicted within a 300 residue segment in the middle of the protein (Figure 5). A β-turn consists of a sequence of four residues that create a bend of nearly 180° in the polypeptide backbone. This is the major structural feature that confers globular shapes to proteins (Chou and Fasman, 1979). The methods of Robson and coworkers and Chou and Fasman predict, respectively, that only 4% and 10% of the residues are involved in β-turn formation within a 315 residue segment in the middle of the 50 kd keratin (Figure 5; 47–362). In contrast, for globular proteins the average frequency of β-turns is 32% (Chou and Fasman, 1979).

Discussion

What Are the Evolutionary Relationships between the 50 kd Epidermal Keratin and Other Intermediate Filament Proteins, and Keratins?

Recent nucleic acid hybridization studies indicate that the cultured human epidermal cell keratin mRNAs can be grouped into two classes, one coding for 56–58 kd keratins and the other for 46–50 kd keratins (Fuchs et al., 1981). We have presented the nucleotide sequences of two cloned cDNAs that are copies of the

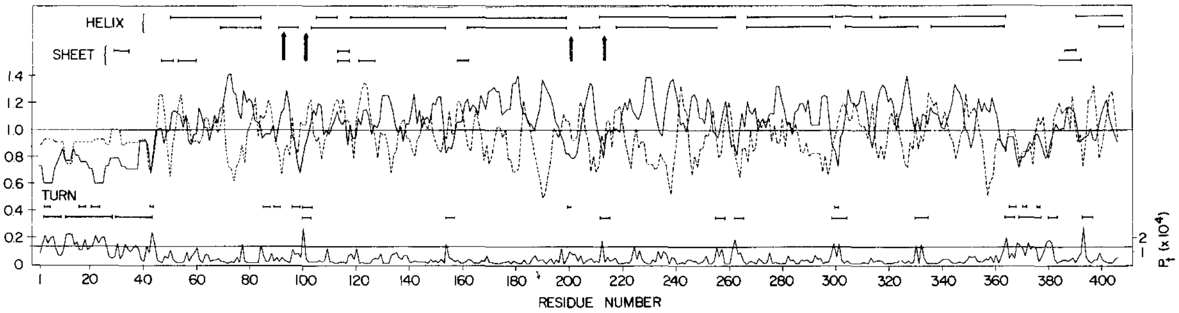

Figure 5. Conformational Profile and Predicted Secondary Structure of the 50 kd Human Epidermal Keratin

The Chou and Fasman (1978, 1979) and Robson and coworkers (see Experimental Procedures) methods used make predictions for four conformational states: α-helix, β-sheet, β-turn and random coil. The predictions for the first three states are shown by the contiguous lines delineated with a vertical bar. In each case, the top line represents the predictions of the Robson method and the bottom line the prediction of the Chou and Fasman (1978, 1979) method. The regions of the protein with no prediction shown are predicted to be in random coil conformation. The residue numbers correspond to the positions in Figure 2. Vertical arrows: positions of prolines and a cluster of glycine that inhibit helical conformation. Left ordinate: average helical potential (——) and sheet potential (- - - -) values for successive tetrapeptides. These values are based on the normalized frequencies of amino acid occurrence in the respective conformational states within proteins of known structure (Chou and Fasman, 1979). Right ordinate: relative probability of turns (Chou and Fasman, 1979). For the Chou and Fasman method, the cutoff point for turns was taken as 1.5(p_t), that is, 1.27 × 10^{-4} (Chou and Fasman, 1979). For the Robson method, the helical potential was assigned a decision constant of −75 as suggested by Garnier et al. (1978).

mRNA for the 50 kd epidermal keratin. At present, we do not know the sequence of a keratin of the 56–58 kd class, but we expect some homology between the two sequences, because most, if not all, cytoskeletal keratins are similar in their amino acid compositions and structural properties (Steinert and Idler, 1975; Fuchs and Green, 1978).

The sequence of the 50 kd epidermal keratin clearly shares homology with two different wool microfibrillar keratin partial sequences (Figure 3), even though the 50 kd keratin is present in a variety of epithelial cells that do not produce hair or wool (Sun and Green, 1978; Bowden and Cunliffe, 1981; Franke et al., 1979). Thus our findings provide strong evidence for an evolutionary relationship between the cytoskeletal keratins and the α-helical microfibrillar keratins. It is important to note, however, that there is a substantial difference in the degree of homology between the 50 kd keratin and each of the two wool keratin fragments (59% versus 27%; Figure 3). These "type-I" and "type-II" wool keratin fragments show only 30% homology with each other. It is not known whether these two fragments generated by partial chymotryptic digestion of microfibrillar keratins originate from the same or different polypeptides. This raises the intriguing possibility that "type-II" wool keratin fragments may be more closely related to the 56–58 kd keratins and "type-I" to the 46–50 kd keratins.

The amino acid sequence of the 50 kd keratin also shares some similarities with those of other intermediate filament proteins (Figure 4). However, the degrees of homology observed among a desmin, a vimentin, a neurofilament protein and the 50 kd keratin indicate that the last represents the most distantly related member of this class of fibrous proteins.

In contrast to the above comparisons, neither the amino acid composition nor the sequence of the 50 kd keratin showed any apparent similarity to those of keratins that constitute the matrix of epidermal appendages or to feather keratins. The sequences of some matrix keratins are unusual in that they contain a series of short (10 residues) tandemly repeating sequences (Barker et al., 1978). The existence of repeated sequences in the cytoskeletal keratins has been postulated to explain the wide variation in their sizes (40–70 kd), despite the close similarity in their amino compositions (Steinert and Idler, 1975; Fuchs and Green, 1978). The glycine-rich amino-terminal region of the protein did indeed contain short repeating sequences (Figure 2); however, outside of this region, tandemly repeating sequences were not observed. Although there were several stretches of 10–30 nucleotides within the KB-2 cDNA that are significantly homologous, these do not all code for a similar amino acid sequence. The evolutionary origins and significance of these duplicated sequences may be elucidated after the coding sequences for different cytoskeletal and microfibrillar keratins from diverse sources are determined.

What Are the Common Structural Features of Intermediate Filament Proteins?

In the absence of knowledge of the complete primary structure of an intermediate filament protein, two major approaches have been taken to elucidate the structural arrangement of the polypeptide components within the filaments: X-ray crystallography, and analysis of the polypeptide components of intact filaments or those of their partial proteolytic fragments (Crick, 1953; Pauling and Corey, 1953; Crewther and Harrap, 1967; Skerrow et al., 1973; Fraser et al., 1976; McLachlan, 1978; Steinert, 1978; Steinert et al., 1980). These studies have indicated that the proteins that constitute the filaments contain two helical domains each about 100–120 residues in length and separated by nonhelical regions, and that these proteins assemble into rope-like coiled-coil structures that are 7–8 nm in diameter.

Computer analysis of the amino acid sequence of the 50 kd keratin predicts a secondary structure that is, in its general outlines, consistent with the physicochemical studies cited above (see Results and Figure 5). However, the previous models, which picture only two long helical regions separated by and tailed with nonhelical regions, may be simplified representations of the structures of these proteins. As the predictions indicate, smaller helical regions may be present outside of the major helical domains, and the nonhelical portions may have a more complex structure than a simple random coil (Figure 4). The analyses on which the previous models were based would not be sensitive enough to identify such regions.

When we conducted similar analyses on the partial sequences of microfibrillar wool keratin (Gough et al., 1978) and porcine desmin and vimentin (Geisler and Weber, 1981; Geisler et al., 1982), we found that despite extensive amino acid sequence differences between these homologous proteins, the sequence of each one was compatible with the formation of long stretches of helical conformation. In the microfibrillar keratin fragments, there appeared to be a periodicity in the occurrence of nonpolar residues similar to that observed in tropomyosin, and it has been suggested that these may constitute a line of hydrophobic residues along one side of the coiled molecule that corresponds to interstrand or interrope contact regions (Fraser et al., 1976; Gough et al., 1978). An examination of our sequence indicated that there is a similar periodicity within the 50 kd segment homologous to the microfibrillar keratin fragments. The residues at these positions appeared to be more strongly conserved than the rest of the sequence.

Our results, together with the results of others (Geisler and Weber, 1981; Dodemont et al., 1982) indicate that there is considerable sequence divergence among intermediate filament proteins. This degree of diversity is markedly different from that of some other structural proteins, for example, collagens and actins, the sequences of which are highly conserved not only between different cellular forms but also across species (for example, Vandekerckhove and Weber, 1978; Bornstein and Sage, 1980; Fuchs et al., 1982). Our comparison of the 50 kd keratin and the microfibrillar keratins suggests that amino acid substitutions in the sequence of intermediate filament proteins can be tolerated without grossly perturbing the helical conformation of a region of the protein that is necessary for their assembly into filaments.

It has been suggested that a 20–22 nm repeat visualized in the fine ultrastructure of all intermediate filaments corresponds to the major helical domains of the protein subunits (Milam and Erickson, 1981; Henderson et al., 1982). If this is correct, then the major variability among the intermediate filament proteins will be observed in the nonhelical segments. We do not yet know how the nonhelical domains are organized in the rope-like coiled-coil; however, variations in the length or sequence of these regions may confer subtle structural differences on intermediate filaments that could be important in meeting the cytoskeletal requirements of diverse cell types. The determination of the complete primary structures of other intermediate filament proteins and the examination of different structural models on the basis of these sequences should help to elucidate some of the many questions posed by their diversity.

Experimental Procedures

Materials

Restriction endonucleases (Bethesda Research Laboratories and New England BioLabs), calf intestinal alkaline phosphatase and T4 polynucleotide kinase (Boehringer and Bethesda Research Laboratories) and γ-^{32}P-ATP and α-^{32}P-dNTPs (Amersham) were purchased from the sources indicated. Reverse transcriptase was obtained from J. Beard (Life Sciences, Inc.).

Preparation of DNA Fragments

Large-scale plasmid purification was carried out with the procedure of Birnboim and Doly (1979), with some modifications (Fuchs et al., 1981). Restriction fragments of DNA were purified from polyacrylamide gels by electrophoresis of the DNA onto Whatman 3MM paper. The procedure used was essentially as described by Girvitz et al. (1980), with the following modifications: We cut a strip of polyacrylamide gel containing the DNA fragment, placed it on a horizontal gel electrophoresis apparatus and embedded it in an agarose gel by pouring a solution of 1.4% agarose, 50 mM Tris–borate (pH 8.3) and 1 mM EDTA at 60°C. We made a slit directly in front of the gel slice, and inserted an L-shaped piece of dialysis membrane along the slit and under the gel to prevent DNA loss during electrophoresis. We then placed a strip of 3MM paper between the dialysis membrane and the gel slice. Electrophoretic contact was maintained with 3MM paper wicks, and the progress of the electrophoresis was monitored by the migration of bromophenol blue placed in a hole in the gel (ethidium bromide staining was not used, to avoid nicking the DNA). For the elution of DNA, we shredded the 3MM paper and placed it with the dialysis tubing in a tube with a microfilter holder (Bioanalytical Systems), which contained 0.3 ml elution buffer (250 mM NaCl, 0.1 mM EDTA and 10 mM Tris [pH 7.4]). After 15 min, we attached an Eppendorf tube to the outlet of the microfilter holder transferred the assembly to a Corex tube and centrifuged it at 2000 rpm for 2 min. After a 0.2 ml wash with elution buffer collected in the same Eppendorf tube, the DNA was precipitated by the addition of 1 ml 100% ethanol at −20°C.

Restriction Site Mapping

Both of the cloned cDNAs sequenced were inserted at the Pst I site of the E. coli plasmid pBR322 (map site 3612; Sutcliffe, 1979). The locations of the internal Pst I sites in the cDNA insert were determined by partial Pst I digestion of a 5′-end-labeled fragment (from Ava I [1424] to Pvu I [3737] of pBR322), as previously described by Smith and Birnstiel (1976). Other sites convenient for labeling and sequencing were located initially by multiple enzyme digests (Danna, 1980) and later by computer searches for restriction sites within fragments already sequenced.

DNA Sequence Analysis

Initially we labeled restriction fragments at their 5′ ends with polynucleotide kinase (Maxam and Gilbert, 1980). Later we used 3′ end labeling for the 5′ end extended restriction cuts. For this procedure, the DNA was incubated in 50 mM Tris (pH 8.1), 40 mM KCl, 7 mM $MgCl_2$ and 4 mM DTT, with 125 μCi of the appropriate α-^{32}P-dNTP and 8 units of reverse transcriptase (as a DNA-dependent DNA polymerase), at 37°C for 30 min. The labeled fragments were either denatured in 30% dimethylsulfoxide or 98% formamide and then electrophoresed to separate the labeled strands, or digested with a second restriction enzyme to obtain fragments labeled at only one end. Purified labeled single-stranded DNAs were then chemically modified and cleaved for DNA sequence analysis (Maxam and Gilbert, 1980). The reaction products were subjected to electrophoresis on 40 cm 10%–12.5% or 85 cm 6% polyacrylamide sequencing gels at 60–70 W.

Amino Acid Analysis of the 50 kd Keratin

Keratins were isolated from cultured human epidermal cells, and individual polypeptides were separated and purified by polyacrylamide gel electrophoresis and electroelution (Fuchs and Green, 1978, 1981). One preparation was carried out with glycine in the gels and buffers, and another by the substitution of molar equivalents of alanine for glycine. Electroeluted samples were dialyzed against 10 mM Tris–HCl (pH 7.6) prior to precipitation with trichloroacetic acid. Aliquots of 20 μg of purified 50 kd keratin were hydrolyzed in 6 M HCl at 108°C for 36 hr in evacuated sealed tubes. After vacuum drying, samples were applied to a Dionex D-501 amino acid analyzer.

Computerized Predictions of Protein Secondary Structure

To predict the secondary structure of the 50 kd keratin, we initially obtained one computer program for the Chou and Fasman (1978, 1979) method from G. Long; however, this did not include prediction of β-turns and it only calculated average conformational parameters for successive tetrapeptides. Therefore, we wrote a program for this method basing our algorithms on their published predictions. As a test, we applied our programs to sequences for which secondary structures had already been predicted by Chou and Fasman (1978). Our computerized predictions of β-turns were identical to theirs, and helix and sheet predictions were very similar, with small differences generally located at the terminal parts of each region. The second method we used was based on that of Garnier et al. (1978). For this method, we received a listing of a computer program (written by B. Robson and coworkers) from B. Robson, which we adapted for our IBM Personal Computer.

Acknowledgments

We are grateful to Naoko Tanese for her very capable technical assistance. In addition, we are grateful to Dr. Raymond J. MacDonald at The University of Texas, Dallas, and Dr. Gerald R. Crabtree at N.C.I., who provided us with protocols on various aspects of DNA sequencing; Dr. George L. Long at University of Washington and Dr. Barry Robson at University of Manchester, who sent us the computer program listings mentioned in the Experimental Procedures; and Mr. George Gibson, who designed and constructed the DNA sequencing apparatuses. Many other friends and colleagues gave us valuable advise on DNA sequencing and we thank them. We also gratefully acknowledge that for DNA sequence analysis we used the Stanford University Molgen Project resources and the N.I.H.-Sumex Aim Facility. I. H. is the recipient of an NIH postdoctoral fellowship. E. F. is a Searle Scholar and the recipient of a Research Career Development Award. This work was supported by a grant from NIH.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *"advertisement"* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received July 23, 1982; revised August 27, 1982

References

Baden, H. P., Goldsmith, L. A. and Fleming, B. (1973). The polypeptide composition of epidermal prekeratin. Biochim. Biophys. Acta *317,* 303–311.

Barker, W. C., Ketcham, L. K. and Dayhoff, M. O. (1978). Duplications in protein sequences. In Atlas of Protein Sequence and Structure, *5,* M. O. Dayhoff, ed. (Washington, D. C.: National Biomedical Research Foundation), pp. 359–362.

Birnboim, H. C. and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl. Acids Res. *7,* 1513–1523.

Bornstein, P. and Sage, H. (1980). Structurally distinct collagen types. Ann. Rev. Biochem. *49,* 957–1003.

Bowden, P. E. and Cunliffe, W. J. (1981). Modification of human prekeratin during epidermal differentiation. Biochem. J. *199,* 145–154.

Brutlag, D. L., Clayton, J., Friedland, P. and Kedes, L. H. (1982). SEQ: a nucleotide sequence analysis and recombination system. Nucl. Acids Res. *10,* 279–294.

Brysk, M. M., Gray, R. H. and Bernstein, I. A. (1977). Tonofilament protein from newborn rat epidermis. J. Biol. Chem. *252,* 2127–2133.

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. *26,* 367–383.

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-carboxymethylkerateine 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.

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

Culbertson, V. B. and Freedberg, I. M. (1977). Isolation and characterization of the α-helical proteins from new born rat. Biochim. Biophys. Acta *490,* 178–191.

Danna, K. (1980). Determination of fragment order through partial digests and multiple enzyme digests. Meth. Enzymol. *65,* 449–467.

Dodemont, H. J., Soriano, P., Quax, W. J., Ramaekers, F., Lenstra, A., Groenen, A. M. A., Bernardi, G. and Bloemendal, H. (1982). The genes coding for the cytoskeletal proteins actin and vimentin in warm-blooded vertebrates. EMBO J. *2,* 167–171.

Dopheide, T. A. A. (1973). The primary structure of a protein component 0.62, rich in glycine and aromatic residues, obtained from wool keratin. Eur. J. Biochem. *34,* 120–124.

Fiddes, J. C. and Goodman, H. M. (1980). The cDNA for the β-subunit of human chorionic gonadotropin suggest evolution of a gene by readthrough into the 3′-untranslated region. Nature *286,* 684–687.

Fitzgerald, M. and Shenk, T. (1981). The sequence 5′-AAUAAA-3′ from part of the recognition site for polyadenylation of late SV40 mRNAs. Cell *24,* 251–260.

Franke, W. W., Schmid, E., Osborn, M. and Weber, K. (1978). Different intermediate-sized filaments distinguished by immunofluorescence microscopy. Proc. Nat. Acad. Sci. USA *75,* 5034–5038.

Franke, W. W., Appelhans, B., Schmid, B., Schmid, E., Freudenstein, C., Osborn, M. and Weber, K. (1979). Identification and characterization of epithelial cells in mammalian tissues by immunofluorescence microscopy using antibodies to prekeratin. Differentiation *15,* 7–25.

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

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. and Green, H. (1978). The expression of keratin genes in epidermis and cultured epidermal cells. Cell *15,* 887–897.

Fuchs, E. and Green, H. (1980). Changes in keratin gene expression during terminal differentiation of the keratinocyte. Cell *19,* 1033–1042.

Fuchs, E. and Green, H. (1981). Regulation of terminal differentiation of cultured human keratinocytes by vitamin A. Cell *25,* 617–625.

Fuchs, E. V., Coppock, S. M., Green, H. and Cleveland, D. W. (1981). Two distinct classes of keratin genes and their evolutionary significance. Cell *27,* 75–84.

Fuchs, E., Kim, K. H., Hanukoglu, I. and Tanese, N. (1982). The evolution and complexity of the genes encoding the cytoskeletal proteins of human epidermal cells. In Biochemistry of Normal and Abnormal Epidermal Differentiation, H. Ogawa and I. Bernstein, eds. (Tokyo: University of Tokyo Press).

Gard, D. L. and Lazarides, E. (1980). The synthesis and distribution of desmin and vimentin during myogenesis in vitro. Cell *19,* 263–275.

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. (1981). Comparison of the proteins of two immunologically distinct intermediate-sized filaments by amino acid sequence analysis: desmin and vimentin. Proc. Nat. Acad. Sci. USA *78,* 4120–4123.

Geisler, N., Plessmann, U. and Weber, K. (1982). Related amino acid sequences in neurofilaments and non-neuronal intermediate filaments. Nature *296,* 448–450.

Girvitz, S. C., Bacchetti, S., Rainbow, A. J. and Graham, F. L. (1980). A rapid and efficient procedure for the purification of DNA from agarose gels. Anal. Biochem. *106,* 492–496.

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.

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

Hendy, G. N., Kronenberg, H. M., Potts, J. T. and Rich, A. (1981). Nucleotide sequence of cloned cDNAs encoding human preproparathyroid hormone. Proc. Nat. Acad. Sci. USA *78,* 7365–7369.

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

Lazarides, E. (1980). Intermediate filaments as mechanical integrators of cellular space. Nature *283,* 249–256.

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

Maxam, A. M. and Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. Meth. Enzymol. *65,* 499–560.

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

Milam, L. and Erickson, H. P. (1981). Shadowed keratin and neurofibrils demonstrate a 20 nm periodicity. J. Cell Biol. *91,* 235a.

O'Donnell, I. J. (1973). The complete amino acid sequence of a feather keratin from emu (Dromaius novae-hollandiae). Aust. J. Biol. Sci. *26,* 415–437.

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.

Pruss, R. M., Mirsky, R., Raff, M. C., Thorpe, R., Dowding, A. J. and

Anderton, B. H. (1981). All classes of intermediate filaments share a common antigenic determinant defined by a monoclonal antibody. Cell *27,* 419–428.

Razin, A., Urieli, S., Pollack, Y., Gruenbaum, Y. and Glaser, G. (1980). Studies on the biological role of DNA methylation. IV. Mode of methylation of DNA in coli cells. Nucl. Acids Res. *8,* 1783–1792.

Schmid, E., Osborn, M., Rungger-Brandle, E., Gabbiani, G., Weber, K. and Franke, W. W. (1982). Distribution of vimentin and desmin filaments in smooth muscle tissue of mammalian and avian aorta. Exp. Cell Res. *137,* 329–340.

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

Smith, H. O. and Birnstiel, M. L. (1976). A simple method for DNA restriction site mapping. Nucl. Acids Res. *3,* 2387–2398.

Steinert, P. M. and Idler, W. W. (1975). The polypeptide composition of bovine epidermal α-keratin. Biochem. J. *151,* 603–614.

Steinert, P. M., Idler, W. W. and Zimmerman, S. B. (1976). Self-assembly of bovine epidermal keratin filaments in vitro. J. Mol. Biol. *108,* 547–567.

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

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. Nat. Acad. Sci. USA *77,* 4534–4538.

Sun, T.-T. and Green, H. (1978). Keratin cytoskeletons in epithelial cells of internal organs. Proc. Nat. Acad. Sci. USA *76,* 2813–2817.

Sun, T.-T., Shih, C. and Green, H. (1979). Cultured epithelial cells of cornea, conjunctiva and skin: absence of marked intrinsic divergence of their differentiated states. Nature *269,* 489–493.

Sutcliffe, J. G. (1979). Complete nucleotide sequence of the Escherichia coli plasmid pBR322. Cold Spring Harbor Symp. Quant. Biol. *43,* 77–90.

Swart, L. S. and Haylett, T. (1973). Studies on the high-sulphur proteins of reduced merino wool. Biochem. J. *133,* 641–654.

Valenzuela, P., Quiroga, M., Zaldivar, J., Rutter, W. J., Kirschner, M. W. and Cleveland, D. W. (1981). Nucleotide and corresponding amino acid sequences encoded by α and β tubulin mRNAs. Nature *289,* 650–655.

Vandekerckhove, J. and Weber, K. (1978). Mammalian cytoplasmic actins are the products of at least two genes and differ in primary structure in at least 25 identified positions from skeletal muscle actins. Proc. Nat. Acad. Sci. USA *75,* 1106–1110.

Wain-Hobson, S., Nussinov, R., Brown, R. J. and Sussman, J. L. (1981). Preferential codon usage in genes. Gene *13,* 355–364.

Weaver, C. A., Gordon, D. F. and Kemper, B. (1981). Introduction by molecular cloning of artifactual inverted sequences at the 5′ terminus of the sense strand of bovine parathyroid hormone cDNA. Proc. Nat. Acad. Sci. USA *78,* 4073–4077.