Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 1896–1900, April 1985
Biochemistry

Structure of a gene for the human epidermal 67-kDa keratin

(intermediate filaments/evolutionary conservation)

LESLYE D. JOHNSON*, WILLIAM W. IDLER*, XIAO-MEI ZHOU*, DENNIS R. ROOP†, AND PETER M. STEINERT*‡

*Dermatology Branch and †Laboratory of Cellular Carcinogenesis and Tumor Promotion, National Cancer Institute, Building 10, Room 12N238, National Institutes of Health, Bethesda, MD 20205

Communicated by George E. Palade, November 13, 1984

ABSTRACT We present the structure and nucleotide sequence of a gene encoding the human epidermal 67-kDa keratin. Three genomic clones were isolated from a λ Charon 4A human genomic library by hybridization to a specific cDNA probe. One clone of 12.3 kilobase pairs was shown by R-loop, DNA sequence, and primer-extension analyses to encode an entire gene of about 6.25 kilobase pairs. Of eight identified introns, seven are located within the region that encodes the central coiled-coil α-helical domain of the protein. Except for one intron located at the end of the region encoding this domain, these do not delineate apparent structural subdomains. The positions of five of the introns exactly coincide with the positions of introns previously reported in the hamster gene for the intermediate filament protein vimentin [Quax, W., Egberts, W. V., Hendricks, W., Quax-Jeuken, Y. & Bloemandal, H. (1983) *Cell* 35, 215–233]. These findings suggest that the human 67-kDa keratin and vimentin genes arose from a common ancestral gene.

The cytoskeleton of epithelial cells contains a class of intermediate filaments (IF) called keratin (or cytokeratin) filaments. Detailed two-dimensional gel electrophoretic analyses of extracts of various epithelial tissues have revealed that a mammalian species expresses a total of about 20 different keratin subunits, which can be divided into distinct acidic (pI 4.5–5.5) and neutral–basic (pI 6.5–7.5) groups (1). This large number of subunits appears to originate at the gene level, since translation of poly(A)-enriched RNA extracted from these tissues reveals that a discrete mRNA species exists for each subunit (2–5). These protein chemical and mRNA translation experiments, as well as analyses involving monoclonal antibodies (6, 7), show that any one epithelial-cell type usually expresses only a few of these subunits, and these always occur as coordinated doublets consisting of at least one acidic and one neutral–basic keratin. For example, the hyperproliferating basal cells of newborn mouse or human epidermis express four keratin subunits: two acidic subunits (50 and 55 kDa in mouse, 50 and 52 kDa in human) and two neutral–basic subunits (59 and 60 kDa in mouse, 56 and 58 kDa in human). Suprabasal epidermal cells, committed to terminal differentiation, express one acidic subunit (59 kDa in mouse or 56.5 kDa in human) and one basic subunit of 67 kDa. This co-expression phenomenon appears to pertain throughout vertebrate epithelia (8). These findings all imply that two keratin subunits, one of each charge group, are required for keratin IF assembly *in vivo*, as was found in earlier *in vitro* assembly experiments (9).

Partial or complete amino acid sequences are now available for several of the epidermal keratins (10–14, 38), and the data show that all of them are constructed on a common plan: a central α-helical domain of conserved length and secondary structure, flanked by end-domains of highly variable size (<50 to >150 residues) and amino acid sequence. Comparisons of the sequences of the α-helical domains of different keratin subunits show that they can be divided into two distinct homologous types: the acidic keratins contain type I α-helical sequences, and neutral–basic keratins contain type II α-helical sequences (refs. 11, 13, and 14, and unpublished observations). The reason for the expression of many different subunits and the structural and functional significance of the coincident expression of two types of keratin subunits in one tissue is not known.

In an interesting counterpoint to the complexity of the keratins of epithelial tissues, the IF proteins vimentin, desmin, and glial fibrillary acidic protein, normally expressed in mesenchymal, muscle, and astroglial tissues, respectively, each consist of a single protein; that is, they form homopolymer IF (9). Their amino acid sequences are known (15–17) and their secondary structures conform to that found for keratin IF subunits, but their α-helical coiled-coil sequences form a distinctly different type III group (18).

In terms of the gene structure of IF subunits, mammalian and avian vimentin are known to be each encoded by a single gene (15, 19), and the structure of the hamster vimentin gene has been reported (15). However, no information is currently available on the complexity or structure of the genes that encode the keratins. In this report, we describe the structure and sequence of a gene for the human epidermal keratin of 67 kDa. This gene system is of particular interest because it seems that the expression of the keratin can be modulated by a variety of extrinsic or intrinsic factors in normal and abnormal epidermis and in cell culture (20, 21).

EXPERIMENTAL PROCEDURES

Isolation of Genomic Clones. cDNA clone pK456 [1.95 kilobase pairs (kbp)] encodes about 75% of the mRNA for the human 67-kDa keratin (38). For the detection of genomic clones, two fragments from pK456 were used: a *Taq* I–*Pst* I fragment (0.4 kbp) corresponding to the 3′-nontranslated region of the mRNA; and a *Pst* I–*Pst* I fragment (0.496 kbp) corresponding to a region encoding part of the α-helix of the protein. These fragments were labeled by nick-translation (22). The λ Charon 4A genomic library originating from a partial *Hae* III digest of human fetal liver DNA (23) was screened with these probes by using *in situ* plaque-hybridization procedures (24, 25). After three rounds of subcloning had identified single-phage plaques, positive clones were amplified for isolation of DNA (25).

Restriction Enzyme Mapping. Restriction endonucleases (New England Biolabs or Bethesda Research Laboratories) were used at about 5 units/μg of DNA. Fragments generated from *Pst* I or *Ava* II digests of the cDNA clone were used as probes for mapping. Genomic clones were digested with

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

Abbreviations: IF, intermediate filament(s); bp, base pair(s); kbp, kilobase pairs; kb, kilobases.
‡To whom reprint requests should be addressed.

either single or multiple restriction enzymes, and the fragments were separated on 0.6–1.2% agarose gels. After transfer to nitrocellulose (26), specific fragments were identified by hybridization to the various cDNA mapping probes. Fragments were sized by comparison with the relative mobilities of digests of λ and ϕX-174 DNA (Bethesda Research Laboratories).

R-Loop and D-Loop Analyses. Cloned genomic DNAs were denatured and then hybridized with poly(A)-enriched RNA (mRNA) isolated from newborn human foreskins (38) or with other genomic DNAs or the cDNA probe for R-loop (27, 28) and D-loop (29) analyses. Relaxed pBR322 circular DNA was included as a standard for double-stranded DNA length measurements.

DNA Sequencing. Only the genomic clone c55 (12.3-kbp *Eco*RI fragment) was sequenced. It was cut once by *Bam*HI, and the two fragments, 8.1 kbp and 4.2 kbp long (see Fig. 2), were inserted into the 4.0-kbp *Bam*HI–*Eco*RI portion of pBR322 and subcloned. The subclones were separated and characterized, and the DNAs were harvested (30). Sequencing was performed using the Maxam–Gilbert (31) procedures following 3′- or 5′-labeling with [^{32}P]phosphate.

Primer Extension Analysis. A *Kpn* I–*Fnu*4HI fragment was identified, labeled at its 5′ ends, and utilized as a primer to extend its mRNA complement (32). The extended fragment then was sequenced directly.

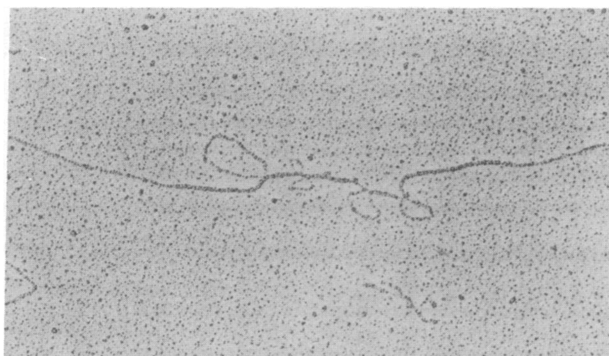
Table 1. Size of exons and introns of the gene

	Size, bp	
	R-loop analysis*	DNA sequence
"CAT box"–"cap site"	—	85
Exon 1	685 ± 80	639
Intron 1	1270 ± 80	≈1170
Exon 2	250 ± 45	215
Intron 2	330 ± 70	318
Exon 3	85 ± 20	60
Intron 3	520 ± 115	484
Exon 4	170 ± 70	97
Intron 4	90 ± 50	167
Exon 5	220 ± 55	173
Intron 5	105 ± 30	113
Exon 6	150 ± 45	117
Intron 6	700 ± 70	≈710
Exon 7	270 ± 45	218
Intron 7	695 ± 40	705
Exon 8	60 ± 45	38
Intron 8	54 ± 30	92
Exon 9	930 ± 95	≈890†
Total	6500 ± 180	≈6250

*Average ± SD of measurements of 10 R-loop structures.
†This number consists of 870 bp plus about 20 bp to the expected start of the poly(A) tail.

RESULTS AND DISCUSSION

Isolation of Genomic Clones. We have described the characterization of a cDNA clone pK456 (about 1.95 kbp) that contains information that encodes about 75% of the human 67-kDa keratin (38). From its deduced amino acid sequence, we found that it lacks the information at the 5′ end of the keratin mRNA, and although it contained a polyadenylylation consensus signal sequence at its 3′ end, it lacked the poly(A) sequence itself. RNA blot–hybridization analyses revealed that the full-length mRNA for this keratin is 2.5 ± 0.1 kilobases (kb). Assuming this length includes a poly(A) tail of about 100 base pairs (bp), the cDNA lacks about 0.5 kbp of information from its 5′ end. Using this cDNA or smaller restriction endonuclease fragments derived from it as probes, a total of three genomic clones were isolated from a λ Charon 4A human genomic library. Each of these was extensively mapped by restriction enzyme digestions and hybridization analyses with the cDNA clone to determine their relatedness and information content. The genomic clones c55 and c509 were each about 12.3 kbp long, gave identical restriction enzyme maps, and hybridized to the full length of the cDNA. The genomic clone c433 (7.3 kbp) appeared to be a shorter subset of the others, as its information did not cover the 5′ end of the cDNA.



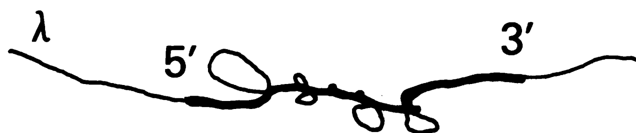


FIG. 1. R-loop analysis of the genomic clone c55 for the 67-kDa keratin gene. By this method, eight loops of intron sequences can be identified over a total estimated length of about 6.5 kbp. (*Upper*) Electron micrograph. (×11,200.) (*Lower*) Representation shows the 5′ and 3′ ends of the gene within the λ clone.

R-Loop Analysis. Fig. 1 shows a typical structure formed between human foreskin epidermal total mRNA, of which that encoding the 67-kDa keratin is a major component, and the genomic clone c55. Eight loops indicating intron sequences are visible. The genomic clone c509 gave identical structures. The orientation of the gene information within the genomic clones was determined by the mapping analyses used above, D-loop analysis with the cDNA (data not shown), and the sequence information described below. From these data, we found that the largest intron is located toward the 5′ end (Fig. 1). The average sizes of the intron and exon regions are summarized in Table 1. The data show

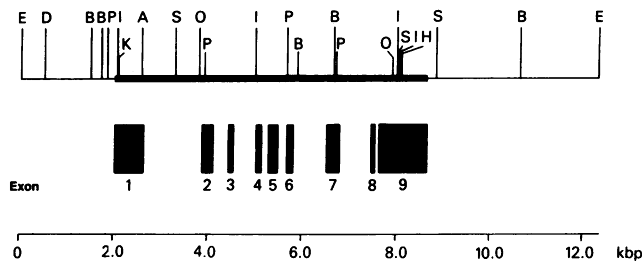


FIG. 2. Physical map of genomic clone c55 (12.3 kbp) and gene (6.25 kbp). The locations of prominent restriction endonuclease sites are shown: A, *Ava* I; B, *Bgl* II; D, *Nde* I; E, *Eco*RI; H, *Bam*HI; I, *Nci* I; K, *Kpn* I; O, *Nco* I; P, *Pst* I; S, *Sac* I. The region encoding the gene is shown by the heavy line. The nine exons representing 2.44 kb of the mRNA are shown by the black boxes.

that the genomic c55 contains the full-length gene, whose estimated total length is about 6.5 kbp.

Nucleotide Sequence and Structure of Gene. Fig. 2 shows a physical restriction endonuclease map of the genomic c55 and Fig. 3 shows the nucleotide sequence of all exons and most introns of the gene. The 3′ end of the gene probably terminates about 20 bp downstream from a polyadenylylation signal sequence in the cDNA sequence (38). Identification of the probable 5′ end of the gene including possible signal sequences for the initiation of transcription and translation was more difficult because complete cDNA information was unavailable. The R-loop analysis (Fig. 1 and Table 1) revealed an exon of about 700 bp at the 5′ end and upstream from a large intron. The DNA sequence shows (Fig. 3) that this intron begins about 150 bp from the 5′ end of the cDNA clone. We then sequenced about 1.3 kbp upstream from this intron to find possible regulatory and transcriptional control consensus sequences and the likely 5′ end of the mRNA. Within this 1.3-kbp region, several potential sites for a "CAT box," a "TATA box," and a "cap site," all of which are commonly found in eukaryotic genes (34, 35), were identified. In addition, extension of the codon reading frame we had found in the cDNA sequence (38) encountered a termination codon just upstream of a *Kpn* I site (Fig. 3). Therefore, to resolve the question of the likely 5′ ends of the gene and mRNA, a 72-bp *Kpn* I–*Fnu*4HI fragment of genomic DNA was isolated, ^{32}P-labeled at its 5′ends, and used as a primer for a primer-extension experiment with mRNA. The primer was extended uniformly by 70 bp, and sequencing of it revealed that the 5′ end of the mRNA continues uninterrupted to the probable cap site indicated in Fig. 3. We conclude that the probable TATA box utilized is located 42 bp upstream from this, and we suggest the possible CAT box is located 85 bp upstream from the cap site. The extension experiment indicates that of all the potential signal sequences identified, only these are used for this gene. The total length of the gene for the 67-kDa keratin is therefore about 6.25 kbp, and the mRNA transcribed from it is 2.44 kb long [excluding the poly(A) tail] (Fig. 3 and Table 1).

Complete Amino Acid Sequence of the 67-kDa Keratin. The definitive location of the cap site of the mRNA for the gene readily permits identification of the initiation (ATG) codon for translation (Fig. 3), since there is only one such codon downstream from the cap site located within the open reading frame recognized from the sequence of the cDNA clone. The deduced amino acid sequence of 643 amino acids is shown above the nucleotide sequence in Fig. 3. The first amino acid is serine, as previously predicted for epidermal keratins (36). The calculated molecular mass of the protein is

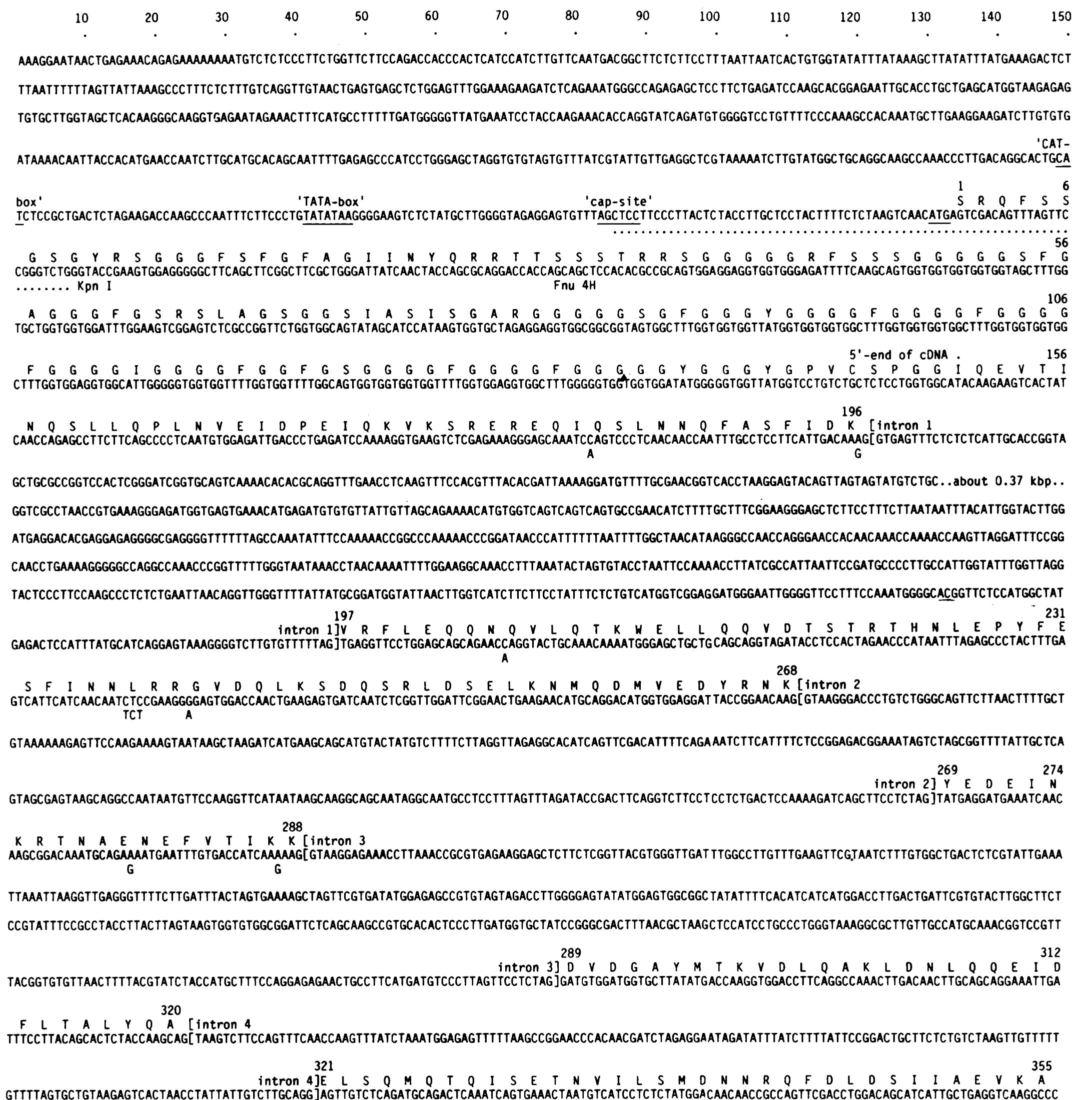


FIG. 3. (*Figure continues on the next page.*)

66.4 kDa, and although its phosphate content is not known, this molecular mass closely conforms to that determined by $NaDodSO_4$/PAGE.

The availability of the complete amino acid sequence for this keratin allows us to extend our observations of the sequence information deduced from the cDNA sequence. The protein contains 179 amino acids in its amino-terminal domain and 151 residues in its carboxyl-terminal domain and, thus, is approximately symmetrical about its central α-helical domain. Moreover, the amino-terminal domain reveals a subdomain structure like that of the carboxyl-terminal domain. For the amino-terminal end, the subdomains are (*i*) N, comprising the first 38 residues, which is likely to form a random-coil conformation due to the frequent occurrence of serine and threonine and is highly basic in charge; (*ii*) V1, a very glycine-rich (60%) sequence of 105 residues; and (*iii*) g1, a globular subdomain 36 amino acids immediately adjacent to the central α-helical domain. Further analyses of the complete sequence information will be needed to characterize the likely structural and functional roles of the end-domains of the 67-kDa keratin. Comparisons of the exact amino acid sequences of the end-domains of this keratin with those of other keratins reveal significant differences (38), which are presumably important in defining the function of the keratin IF in the terminally differentiating cells in which this keratin is expressed.

Fig. 3 shows that a total of 36 nucleotides are changed between the available cDNA sequence and this gene sequence, which result in 12 amino acid substitutions. Of these 12, 6 are located in the central α-helical domain, resulting in 4 fewer positive charges, but their locations are such that they are unlikely to significantly change the coiled-coil structure. The origin of the differences between the nucleotide sequences is not clear. We speculate that these substitutions may have arisen as a result of reverse transcriptase errors in copying the mRNA sequence or as a result of λ phage amplification during preparation of the genomic clones. These substitutions may also represent polydispersity in the mRNA from which the cDNA library was constructed, since the mRNA was isolated from the pooled epidermal tissue of many individual foreskins. We cannot, however, exclude the possibility that a family of similar genes exist for the 67-kDa keratin and that more than one of these is expressed in the epidermis. Preliminary data from our laboratory suggests that there may be six or more copies of the gene in the human haploid genome.

Location of Introns in the 67-kDa Keratin Gene. To our knowledge, the only IF protein gene that has been characterized is that for hamster vimentin (15), but despite the paucity of information, it is nevertheless interesting to compare the exon/intron structure of this keratin and that of vimentin. Fig. 4 shows that vimentin has six and the keratin, seven

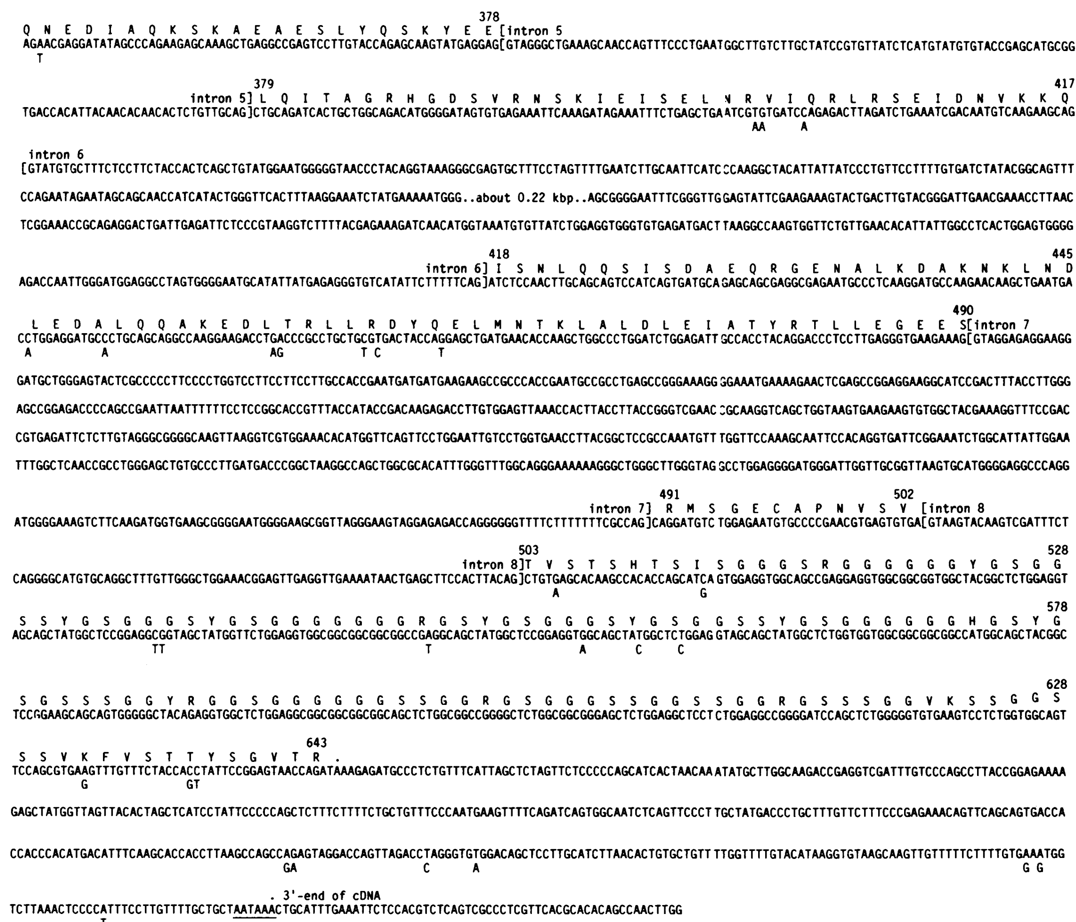


FIG. 3. DNA and amino acid sequences of the gene for the 67-kDa keratin. Regions encoded by the cDNA described elsewhere are indicated. Introns are delineated by square brackets. Introns 1 and 6 were incompletely sequenced, but their sizes were estimated from the mobilities of DNA fragments of known size. The *Kpn* I–*Fnu*4HI fragment used for primer extension is shown, and the sequence extended in the experiment is indicated by the dotted line (nucleotide positions 686–758). The possible "CAT box" and likely "TATA box," "cap site," initiation codon, and polyadenylylation signal sequences are underlined. The occasional nucleotides shown below the line are those that differ from the cDNA clone. The deduced amino acid sequence of 643 amino acids is shown above the nucleotide sequence, using the single-letter code (33).

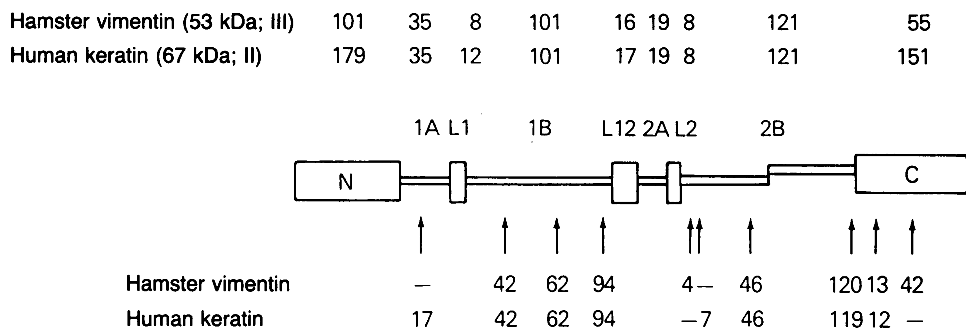


FIG. 4. Comparative structures of the human 67-kDa keratin and hamster vimentin genes. Roman numerals (II and III) refer to α-helix types. The conserved secondary structure of the proteins consists of the end-domains N and C and the central α-helical domain. The latter is subdivided into the 4 coiled-coil subdomains 1A, 1B, 2A, and 2B (boxes) by the non-coiled-coil linkers L1, L12, and L2. The coiled-coil subdomain 2B has a heptad polarity discontinuity (stutter) where shown. Above is shown the number of amino acid residues in each domain or subdomain of the two proteins. Below are arrows that delineate positions of introns in the genes. Numbers below refer to the amino acid number within the subdomain or C end-domain.

introns in the central α-helical domain. Remarkably, four of these are located in exactly the same positions in the two proteins, and two others are only 1–3 amino acids different in location. However, only one common intron, that near the end of the coiled-coil tract 2B, appears to closely delineate a structural domain or subdomain. Presumably the evolutionary driving forces that have respected the common secondary structures of these IF proteins have also preserved the positions of several of the introns in their genes. Similar observations with respect to the positions of introns have been made in other protein systems for which the secondary structures have been rigorously conserved (37). Although the structural and functional significance of the conservation of location of introns in these two IF proteins is not clear, we conclude that vimentin and the 67-kDa keratin have diverged from a common ancestral gene. We anticipate that data on the structures of other IF genes will confirm this concept and perhaps provide clues on the subsequent divergence and evolution of the putative ancestral gene into the three known helical types of IF subunits.

Note Added in Proof. Kim *et al.* (39) have suggested that the mRNA for the human 67-kDa type II keratin is 3.8 kb long, based on blot analysis with a nonspecific type II cDNA clone. The complete structure of this gene, reported here, shows that its mRNA is actually about 2.5 kb long. This reaffirms the view (4) that nonspecific cDNA clones must be used with caution.

We thank Drs. Carole Heilman and Donald Court for many useful discussions and much assistance during the isolation and characterization of the genomic clones. We also thank Margery Sullivan for the R-loop analyses.

1. Moll, R., Franke, W. W., Schiller, D. L., Geiger, B. & Krepler, R. (1982) *Cell* **31**, 11–24.
2. Fuchs, E. V. & Green, H. (1979) *Cell* **17**, 573–582.
3. Gibbs, P. E. M. & Freedberg, I. M. (1982) *J. Invest. Dermatol.* **74**, 382–388.
4. Roop, D. R., Hawley-Nelson, P., Cheng, C. K. & Yuspa, S. H. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 716–720.
5. Magin, T. M., Jorcano, J. L. & Franke, W. W. (1983) *EMBO J.* **2**, 1387–1392.
6. Eichner, R., Bonitz, P. & Sun, T.-T. (1984) *J. Cell Biol.* **98**, 1388–1396.
7. Sun, T.-T., Eichner, R., Tseng, S. C. G. & Weiss, R. A. (1983) *J. Invest. Dermatol.* **81**, 109s–115s.
8. Fuchs, E. V. & Marchuk, D. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 4321–4325.
9. Steinert, P. M., Idler, W. W., Anyardi-Whitman, M. & Goldman, R. D. (1982) *Cold Spring Harbor Symp. Quant. Biol.* **46**, 465–474.
10. Hanukoglu, I. & Fuchs, E. V. (1982) *Cell* **31**, 243–252.
11. Hanukoglu, I. & Fuchs, E. V. (1983) *Cell* **33**, 915–924.
12. Steinert, P. M., Rice, R. H., Roop, D. R., Trus, B. L. & Steven, A. C. (1983) *Nature (London)* **302**, 794–800.
13. Steinert, P. M., Parry, D. A. D., Racoosin, E. L., Idler, W. W., Steven, A. C., Trus, B. L. & Roop, D. R. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5709–5713.
14. Steinert, P. M., Idler, W. W., Zhou, X.-M., Johnson, L. D., Parry, D. A. D., Steven, A. C. & Roop, D. R. (1985) *Ann. N.Y. Acad. Sci.*, in press.
15. Quax, W., Egberts, W. V., Hendricks, W., Quax-Jeuken, Y. & Bloemandal, H. (1983) *Cell* **35**, 215–223.
16. Geisler, N. & Weber, K. (1982) *EMBO J.* **1**, 1649–1656.
17. Lewis, S. A., Balcarek, J. M., Krek, V., Shelanski, M. L. & Cowen, N. J. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2743–2746.
18. Steinert, P. M., Jones, J. C. R. & Goldman, R. D. (1984) *J. Cell Biol.* **99**, 225–275.
19. Zerner, Z. E. & Paterson, B. M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 911–915.
20. Fuchs, E. V. & Green, H. (1981) *Cell* **25**, 617–625.
21. Roop, D. R., Toftgard, R., Kronenberg, M. S., Clark, J. H. & Yuspa, S. H. (1984) *Cold Spring Harbor Symp. Quant. Biol.*, in press.
22. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237–251.
23. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell* **15**, 1157–1174.
24. Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180–182.
25. Hager, G. L., Chang, E. H., Chan, H. W., Garon, C. F., Israel, M. A., Martin, M. A., Scolnick, E. M. & Lowy, D. R. (1979) *J. Virol.* **31**, 795–809.
26. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517.
27. Gross-Bellard, M., Oudet, P. & Chambeau, P. (1973) *Eur. J. Biochem.* **36**, 32–38.
28. Thomas, M. R., White, L. & Davis, R. W. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 2294–2298.
29. Tilghman, S. M., Curtis, P. J., Teimeier, D. C., Leder, P. & Weissman, C. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1309–1313.
30. Davis, R. W., Simon, M. & Davidson, W. (1971) *Methods Enzymol.* **21**, 4113–4128.
31. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
32. Roop, D. R., Kristo, P., Stumph, W. E., Tsai, M. J. & O'Malley, B. W. (1981) *Cell* **23**, 671–680.
33. IUPAC-IUB Commission on Biochemical Nomenclature (1968) *Eur. J. Biochem.* **5**, 151–153.
34. Sadler, J. R., Waterman, M. S. & Smith, T. S. (1983) *Nucleic Acids Res.* **11**, 2221–2231.
35. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeReil, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980) *Cell* **21**, 653–668.
36. Steinert, P. M. & Cantieri, J. C. (1983) in *Biochemistry and Physiology of the Skin*, ed. Goldsmith, L. A. (Oxford Univ. Press, New York), pp. 135–169.
37. Nudell, U., Katcoff, D., Zakut, R., Shani, M., Camon, Y., Finer, M., Czosnek, H., Ginsberg, I. & Yaffe, D. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 2763–2767.
38. Steinert, P. M., Parry, D. A. D., Idler, W. W., Johnson, L. D., Steven, A. C. & Roop, D. R. (1985) *J. Biol. Chem.*, in press.
39. Kim, K. H., Marchuk, D. & Fuchs, E. (1984) *J. Cell Biol.* **99**, 1872–1877.