Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 1114–1118, February 1988
Cell Biology

Sequence of the human 40-kDa keratin reveals an unusual structure with very high sequence identity to the corresponding bovine keratin

(DNA sequence/protein structure/intermediate filaments/evolutionary conservation)

RICHARD L. ECKERT

Department of Physiology and Biophysics, Case Western Reserve University School of Medicine, 2119 Abington Road, Cleveland, OH 44106

Communicated by Oscar D. Ratnoff, October 19, 1987

ABSTRACT **The complete amino acid and DNA sequences of the human 40-kDa keratin are reported. The DNA sequence encodes a protein of 44,098 Da, which is unique in that it lacks the terminal non-α-helical tail segment found in all other keratins. When the human 40-kDa keratin amino acid sequence is compared to the corresponding bovine keratin, the overall identity is 89%. The coil-forming regions are 89% identical and the head regions are 88% identical. This similarity is also evident in the DNA sequence of the coding region, the 5′ upstream sequences, and the 3′ noncoding sequences. The high degree of cross-species identity between bovine and human 40-kDa keratins suggests that there is strong evolutionary pressure to conserve the structure of this keratin. This in turn suggests an important and universal role for this intermediate filament subunit in all species.**

Intermediate filaments (IF) (7–12 nm) are important cytoskeletal elements of most vertebrate cells and are encoded by a large multigene family of proteins (1–3). Biochemical studies and DNA sequence analysis have revealed conserved features of all IF proteins (1–4). These include a central rod domain consisting of ≈310 amino acids, which is capable of formation of coiled-coil α-helices, and amino- and carboxyl-terminal domains, which are non-α-helical. The carboxyl-terminal tail and amino-terminal head regions are variable in length and are largely responsible for the size differences among IF proteins (4).

Each IF protein displays a tissue-specific pattern of expression. Vimentin occurs in cells of mesenchymal origin, desmin is typically found in muscle cells, neurofilaments are found in neuronal cells, and glial filaments are found in astrocytes. The most complex pattern of expression is observed for the keratins, which are specifically expressed in epithelial cells (1, 2, 5, 6). At least 19 different peptides are expressed in human tissues in a cell-type-specific manner (5, 6).

The various subfamilies of IF are only distantly related, sharing 25–30% identity (3, 4). These similarities are usually confined to restricted regions of the α-helical rod, especially the amino- and carboxyl-terminal ends (3, 4). Within a given subfamily the similarities are much higher (50–70%) with the α-helical rod exhibiting the highest degree of similarity (3, 4).

The 40-kDa type I (acidic) keratin is encoded by a single mRNA and is the smallest human keratin (5, 7, 8). Its expression is increased by vitamin A treatment (8–10) of cultured human keratinocytes and it is unique in that it is the only type I (acidic) keratin that does not have a type II (basic) partner (6). This manuscript reports the sequence of the human 40-kDa keratin, which reveals an unusual structure and a very high sequence identity at the DNA and amino acid sequence level to the corresponding bovine keratin.*

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.

MATERIALS AND METHODS

Cloning Procedures. cDNA libraries were prepared with mRNA isolated from human keratinocytes treated with 20 nM arotinoid Ro 13-6298 for 4 days to increase the level of 40-kDa keratin mRNA (9, 10). cDNA was generated using the RNase H procedure (11), fitted with *Eco*RI/*Sma* I adaptors, and cloned into *Eco*RI-digested λgt10 (12). A human genomic library, kindly provided by P. Leder (Harvard Medical School), was used to isolate genomic clones. The libraries were screened at high stringency using plasmid pK19-1 (9) labeled with [^{32}P]dCTP by nick-translation (13). pK19-1 is specific for mRNA encoding keratin 19 (9). Sequencing was by the chemical method (14) using rapid chemical sequencing vector pSP65CS (15) or by the chain-extension method utilizing M13 vectors (16). The complete sequence was determined in both directions.

RESULTS

Human 40-kDa Keratin Nucleotide and Amino Acid Sequence. Fig. 1 shows the complete DNA and amino acid sequences of the human 40 kDa keratin 19 plus the first 264 bases of 5′ upstream and noncoding sequence. The total length of the mRNA is 1338 nucleotides, including 10 residues of the poly(A) tail. Assuming an actual poly(A) length of 200 bases, the mRNA size is ≈1530 nucleotides. This is similar to a previous estimate of 1585 nucleotides (9), and is similar in size to the bovine 40-kDa keratin mRNA (18). The coding sequence includes 1200 nucleotides and encodes 400 amino acids corresponding to a molecular mass of 44,098.

40-kDa Keratin Protein Structure Analysis. Protein conformational analysis (refs. 19 and 20; data not shown) revealed structural features common to other intermediate filament proteins (4, 21). These are outlined in Fig. 1. The protein is composed of three α-helical domains, 1A, 1B, and 2, which consist of repeats of the heptad pattern $(abcdefg)_n$ in which a and d are hydrophobic amino acids. This heptad repeat is charactristic of sequences able to form coiled-coil structures (22). Domains 1A, 1B, and 2 are separated by short regions of sequence having non-α-helical properties. The total length of the α-helical rod domain is 315 amino acids. All intermediate filaments contain a sequence Thr-Tyr-Arg-Xaa-Leu-Leu-Gln-Gly-Glx that is located at the end of coil 2 and precedes the change from α-helical to non-α-helical se-

Abbreviation: IF, intermediate filament(s).

*The sequence reported in this paper is being deposited in the EMBL/GenBank data base (Bolt, Beranek, and Newman Laboratories, Cambridge, MA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J03607).

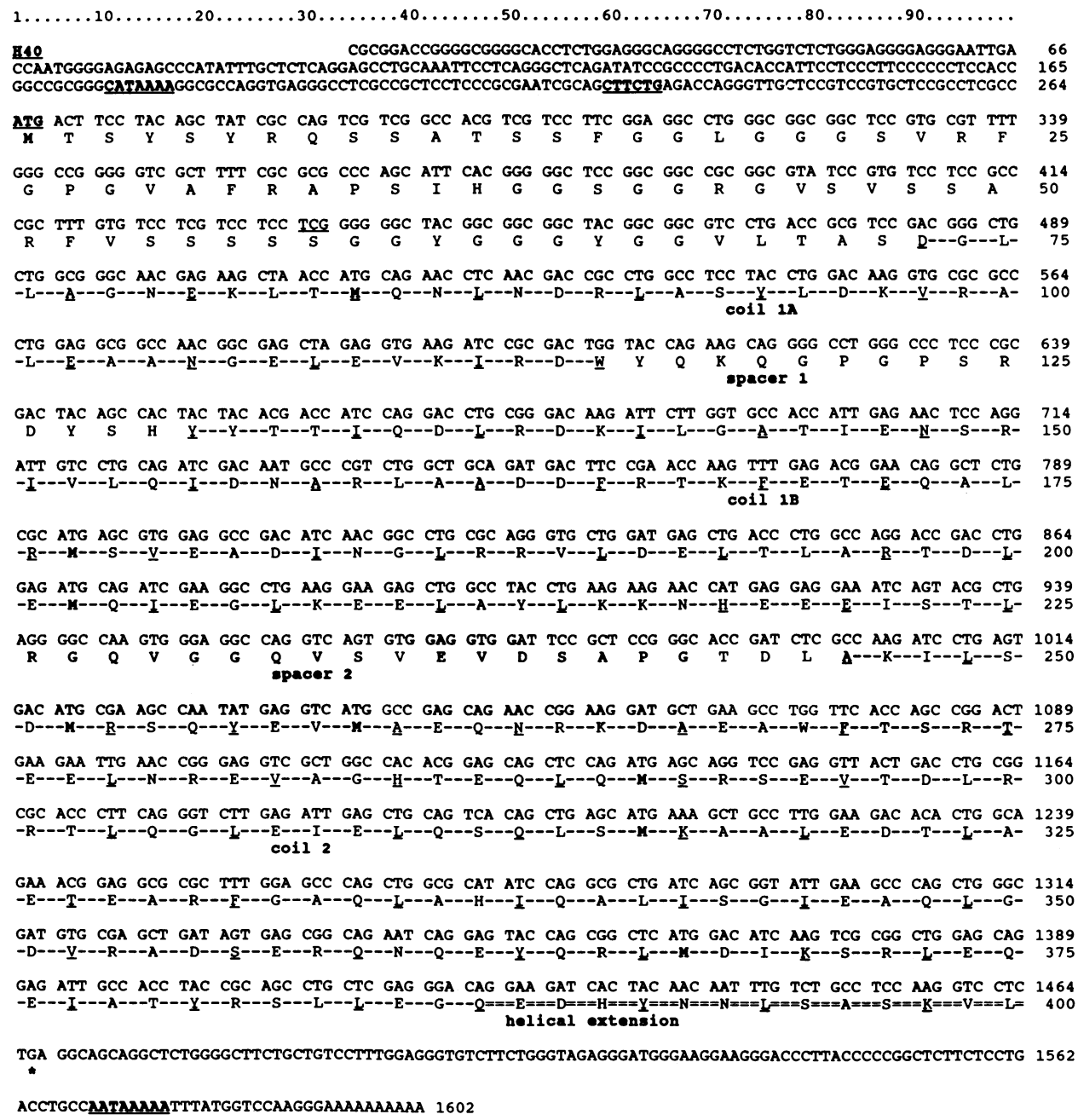

FIG. 1. Nucleotide and amino acid sequences of the 40-kDa human keratin. The nucleotide sequence is presented on the top line and was assembled as follows: cDNA clone pK19-2 included nucleotides 480–1602, and the sequence derived from genomic clone GK19-1 covered nucleotides 1 to the *Kpn* I site at nucleotide 610. Sequence from each clone was identical in the overlapping regions. The amino acids are presented by the single letter designation. The α-helical coiled-coil-forming regions are connected by dashed lines and the hydrophobic amino acid constituents of the heptad repeats are underlined in the rod segment and in the α-helical extension. The 13 amino acids comprising the α-helical extension are joined by double lines. Segments of the α-helical region: coil 1A, coil 1B, coil 2, and spacers 1 and 2 are indicated. The sequence CATAAAA corresponding to the TATA box is underlined and is in bold-face type as is AATAAA, the polyadenylylation signal and the sequence CTTCTG found downstream of the cap site in many eukaryotic mRNAs (17). The translation termination signal is marked by an asterisk.

quence. The 40-kDa keratin also contains this consensus (Fig. 1), but it does not possess a non-α-helical tail; in its place is a 13-amino acid extension of coil 2, which has α-helical character. A similar extension has recently been reported for the bovine 40-kDa keratin (18). The head domain, encompassing amino acid residues 1–72, contains nine basic and no acidic amino acids. It is interesting that the two proteins differ in size by only one amino acid; the human 40-kDa protein has an extra serine residue (position 58; Fig. 2) in a run of five serines. The bovine keratin contains four serines in this series. Serine runs are characteristic of the amino- and carboxyl-terminal non-α-helical segments of keratins (Fig. 2; refs. 3, 18, 23, and 24). The human 40-kDa keratin only contains two regions, amino acids 16–22 and 59–68, that contain the glycine-rich motif, Gly-Gly-Gly-Xaa, which is found repeated in the head and tail segments of many other keratins (3, 23–25).

Amino Acid Sequence Comparison with Other Type I Keratin Polypeptides. Fig. 2 and Table 1 compare the sequence of the human 40-kDa keratin (H40) with the bovine 40-kDa (B40), and the human 46-kDa (H46) and 50-kDa (H50) type I keratins. Overall, H40 is 58–59% identical to H46 and H50, with most of this identity in the α-helical segment. H40 is strikingly less similar to H46 and H50 in the

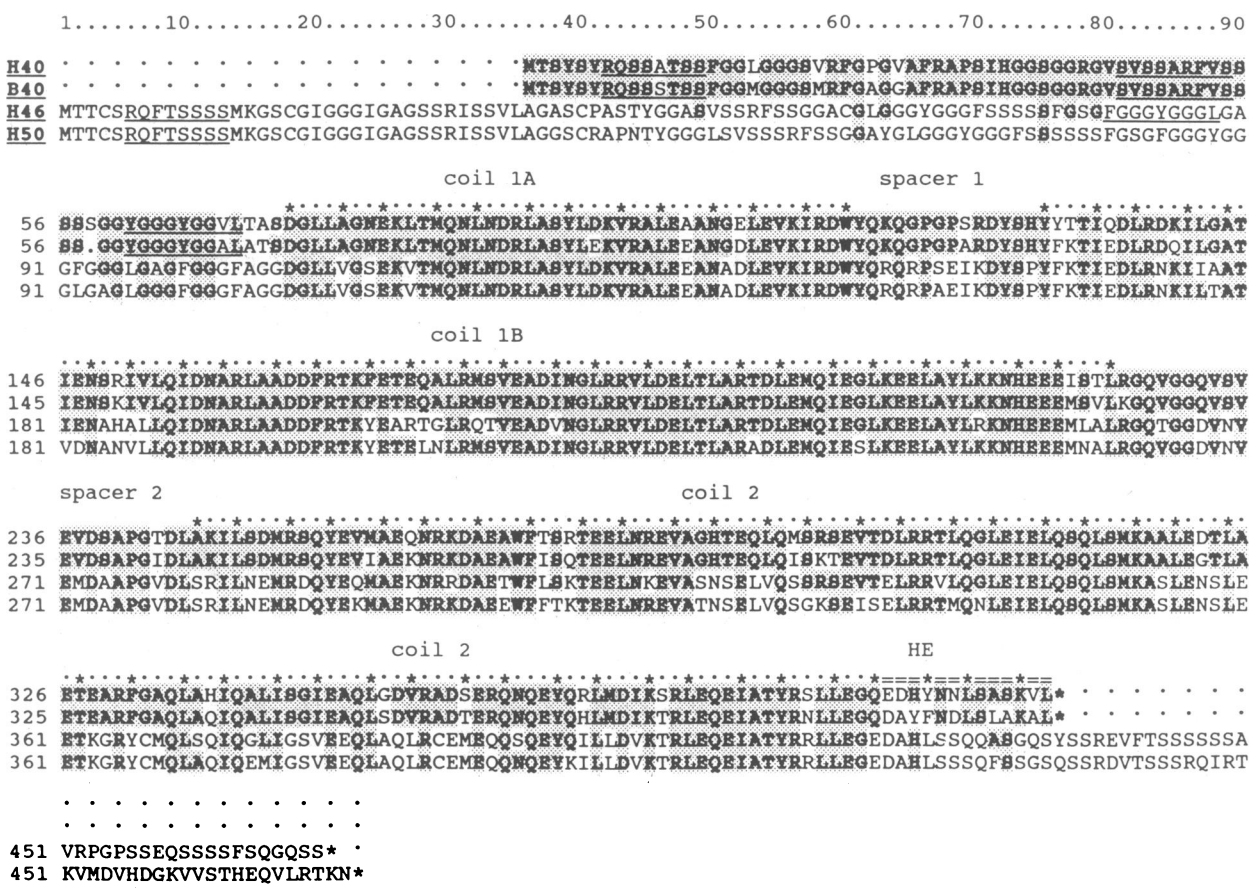

FIG. 2. Amino acid sequence identity comparison of the human 40-kDa keratin with other type I keratins. The amino acid sequences of the human 40- (keratin 19), bovine 40- (keratin 19), human 46- (keratin 17), and human 50- (keratin 14) kDa keratins were aligned for maximal identity. Amino acids are identified by the single-letter code. The α-helical rod includes coil 1A, spacer 1, coil 1B, spacer 2, and coil 2. Hydrophobic amino acids comprising the heptad repeats are marked by asterisks and the extent of the heptads are shown by a dotted line connecting the asterisks. HE, marked by double lines overlying the sequence, designates the 13-amino acid α-helical extension, which replaces the non-α-helical tail in H40 and B40. In H46 and H50, the non-α-helical tail begins with the sequence DAHLSS, etc. The sequence in front of the rod region is the head segment (all keratins). Amino acid positions that are identical to the 40-kDa keratin are indicated by shading. Dotted lines indicate spaces inserted to align the sequences. Notice that the space inserted into the B40 sequence at position 58 accounts for the extra serine in the H40 sequence. The asterisk at the end of each sequence is the translation stop. The sequence information for B40, H46, and H50 is taken from refs. 18, 23, and 24, respectively. The underlined sequences in the head segments are similar to sequences conserved in other keratins.

head region (Table 1). H46 and H50 are more similar to one another than to H40, but the head and tail segments are significantly less similar than the rod segment. H40 and B40 are the most similar; the complete sequence is 89% identical, the coiled-coil α-helical domains are 89% identical, and the head segments are 88% identical. Partial sequence information for the mouse and human 67-kDa keratins is available for a second cross-species comparison (25, 26). H67 and M67 are 69% identical overall. The coil segments are 80% identical, the tail segments are 50% identical, and the available head segment sequences are 64% identical. H40 and B40 are, therefore, very similar compared to other keratin pairs. In contrast to all other keratin pairs, the identity is not confined to the coil segments but extends the entire length of the protein.

DNA Sequence Comparison of Human and Bovine 40-kDa Keratin mRNA. In Fig. 3 the mRNA sequence of H40 and B40 are aligned. Analysis indicates that the two sequences are overall 83% identical, the 5′-noncoding/upstream regions are 61% identical, and the 3′-noncoding regions are 70% identical. The coding regions are 90% identical. This is unusually high similarity when compared to other keratin pairs. For example, the human 46- (keratin 17) and 50- (keratin 14) kDa keratins are 85% identical in the coding region and not significantly similar in the 5′ and 3′ regions (not shown) unless gaps are inserted to align the sequences (23). A cross-species comparison of the human and mouse

Table 1. Amino acid sequence identities among type I keratins

Comparison	Region*	% identity
H40 and B40	Total (1–401)	89
	Head (1–73)	88
	Coil (73–401)	89
	Tail	—
H40 and H46†	Total (1–401)	59
	Head (1–73)	28
	Coil (73–401)	69
	Tail	—
H40 and H50†	Total (1–401)	58
	Head (1–73)	30
	Coil (73–401)	69
	Tail	—
H46 and H50†	Total (1–471)	77
	Head (1–107)	66
	Coil (108–412)	87
	Tail (412–471)	46

*Indicates the amino acid residues from the H40 or H46 keratin sequences selected for use in scanning the B40, H46, and H50 keratin sequences. The amino acid numbers correspond to those in Fig. 2. H40 cannot be compared to H46 and H50 in the tail region, since H40 lacks the non-α-helical tail segment. No gaps were inserted into the sequence to improve fit.

†Sequence of H46 and H50 are from Raychaudhury *et al.* (23) and Marchuk *et al.* (24), respectively. The sequence of B40 is from Bader *et al.* (18).

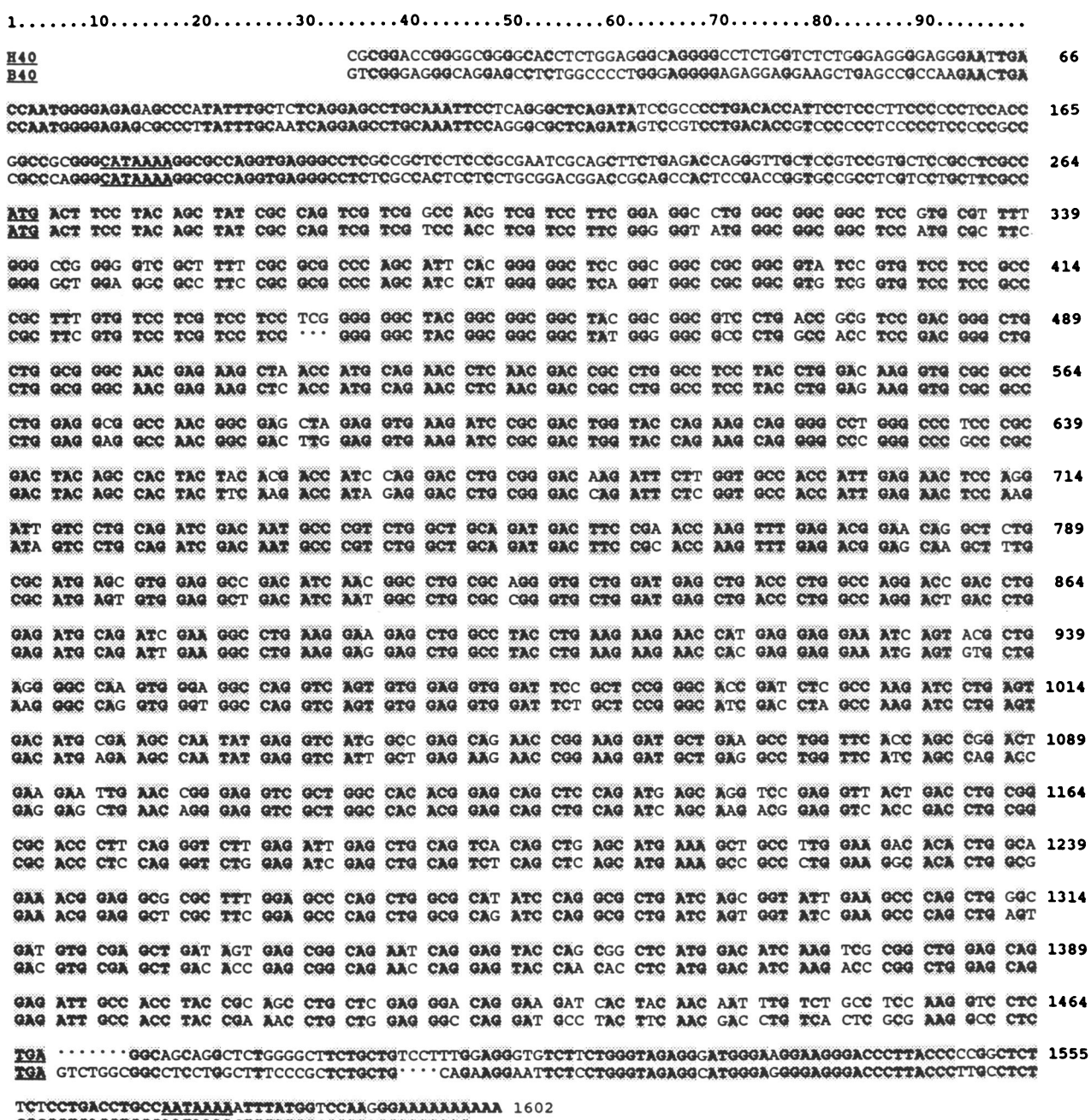

FIG. 3. Nucleotide sequence comparison of the human and bovine 40-kDa keratins. The DNA sequence of the human (H40) and bovine (B40) 40-kDa keratins are aligned for optimal identity. The TATA box sequence is underlined, as is the polyadenylylation signal. The stretch of adenylate residues at the end of each sequence is the poly(A) tail. Identity is indicated by shading. Gaps inserted into the sequence to improve identity are indicated by dotted lines and a gap inserted into the bovine sequence to account for the extra serine in the human sequence is indicated by three dots. The nucleotide position of the human sequence is indicated on the right side.

67-kDa keratins (26), H67 and M67, reveals that the DNA sequences encoding the rod segments are 79% identical, the tail segments are 49% identical, and the noncoding sequences are ≈35% identical (data not shown). Thus, even without extensive gapping to align the sequences, H40 and B40 are extremely similar relative to other keratin comparisons.

DNA Sequence Similarity in the 5′ Regulatory Regions. Sequence similarity in the 5′-noncoding and regulatory regions of H40 and B40 is shown in Fig. 3. Several stretches containing blocks of 9–30 bases are identical, and in a 140-nucleotide stretch (bases 60–200) the identity is 82%. These include a region of strong identity surrounding the "TATA" box sequence (underlined) and a stretch of nucleotides from base 60 to base 130.

DISCUSSION

Amino acid sequence similarity among members of each keratin subfamily (type I acidic; type II neutral–basic) varies; generally the α-helical rod regions are 50–70% identical and the head and tail regions are much less similar (Table 1; refs. 4 and 25). Higher sequence identity has been described for sequences encoding two human keratins of 46 (keratin 17) and 50 (keratin 14) kDa (23). The coding regions of these two human keratin genes are 85% identical, but similarity in the 3′-noncoding sequence is low. This high degree of similarity is perhaps not surprising, since these genes are both human and are of similar size. In addition, in a partial sequence comparison of the bovine and human 50-kDa keratins (27), a stretch of 90 amino acids encompassing the

last 43 amino acids of the coil 2 segment and 50 amino acids of the non-α-helical tail segment of the bovine (keratin VII) and human (keratin 14) 50-kDa keratins were identical with only two exceptions (27).

The data presented herein represent a full-length comparison of corresponding keratins from different species. The sequence comparison of the human and bovine 40-kDa DNA sequences indicates very high similarity even when the sequences are minimally gapped (Fig. 3). The overall DNA sequence identity is 83%. A comparison of the amino acid sequence shows that the head and coil segments, as well as the total sequence, are nearly 90% identical (Table 1; Fig. 2). Thus, H40 and B40 appear to be the most highly conserved intermediate filament pair known. Moreover, the pattern of identity is strikingly different from all other known keratin comparisons in that the non-coil segments are as similar as the coiled segments.

Like B40 (18), H40 is unique in that the non-α-helical tail present in other keratins is replaced by a 13-amino acid α-helical extension. The heptad repeat structure within the extension is in phase with the heptads in coil 2 (Figs. 1 and 2). It therefore appears unlikely that this fragment could have replaced a non-α-helical segment from an ancestral keratin. In fact, the ancestral α-helical sequence may have extended much further. It appears more likely that the 40-kDa is the oldest keratin and that the non-α-helical tail present in all other keratins replaced the α-helical extension prior to generation of other members of the keratin family.

The high degree of cross-species conservation in the head domain, which is not highly conserved between other keratins, is striking and suggests that the conservation is maintained for important functional reasons. Because it is the only unpaired type I keratin (6), the 40-kDa keratin may be a scavenger of type II keratins serving to correct a type I/type II keratin imbalance. This would have obvious advantages for the cell, since it would prevent the appearance of short nonfilamentous structures observed when only one keratin type is present (28, 29). This might be especially important during a change in differentiated status of the cells, such as that promoted by vitamin A (10). A second possible role for this keratin is that of forming filament structures that differ from all others. Although the bovine 40-kDa keratin assembles filaments with appropriate type II partners (18, 28), these filaments may differ in ways that have not yet been detected. These "different" filaments may be important in maintenance of IF structures appropriate to proliferating and/or simple epithelial cells. Because the 40-kDa keratin lacks the carboxyl-terminal tail and because the α-helical extension is short, this end of the molecule is likely to participate in filament formation and/or interact with other cellular structures in a very different way compared to other keratins and to potentially produce alterations in cell morphology and function. This possibility is consistent with the observation that proliferating undifferentiated epidermal cells, in general, produce a large amount of 40-kDa keratin and when the cells differentiate, less of it and more of other type I (and type II) partners is made.

If high overall conservation is a general property of corresponding keratins from different species, it argues that each distinct keratin is conserved for a highly specific, perhaps subtly different, function. This would help to explain the wide diversity in the pattern of expression, size (5, 6), and filament assembly properties (28, 29) of keratins.

The high degree of conservation in the 40-kDa keratin gene 5′- and 3′-noncoding DNA sequence suggests that these sequences are functionally important and that expression of the genes may be regulated in a similar manner in both species.

This work was supported by awards from the Ohio Board of Regents and the Life and Health Insurance Medical Research Fund. I gratefully acknowledge the support and encouragement of Dr. Howard Green in the early phases of this project.

1. Franke, W. W., Schmid, E., Schiller, D. L., Winter, S., Jarasch, E.-D., Moll, R., Denk, H., Jackson, B. W. & Illmensee, K. (1982) *Cold Spring Harbor Symp. Quant. Biol.* **46,** 431–453.
2. Lazarides, E. (1982) *Annu. Rev. Biochem.* **51,** 219–250.
3. Weber, K. & Geisler, N. (1984) in *Cancer Cells 1: The Transformed Phenotype*, eds. Levine, A. J., Vande Woude, G. F., Topp, W. C. & Watson, J. D. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 153–160.
4. Steinert, P. M., Stevens, A. C. & Roop, D. R. (1985) *Cell* **42,** 411–419.
5. Moll, R., Franke, W. W., Schiller, D. L., Geiger, B. & Krepler, R. (1982) *Cell* **31,** 11–24.
6. Sun, T.-T., Eichner, R., Schermer, A., Cooper, D., Nelson, W. G. & Weiss, R. A. (1984) in *Cancer Cells 1: The Transformed Phenotype*, eds. Levine, A. J., Vande Woude, G. F., Topp, W. C. & Watson, J. D. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 169–176.
7. Wu, Y.-J. & Rheinwald, J. G. (1981) *Cell* **25,** 627–635.
8. Fuchs, E. V. & Green, H. (1981) *Cell* **25,** 617–625.
9. Eckert, R. L. & Green, H. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 4321–4325.
10. Gilfix, B. M. & Eckert, R. L. (1985) *J. Biol. Chem.* **260,** 14026–14029.
11. Gubler, U. & Hoffman, B. J. (1983) *Gene* **25,** 263–269.
12. Young, R. A. & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1194–1198.
13. Rigby, P. W. J., Dieckman, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113,** 237–251.
14. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
15. Eckert, R. L. (1987) *Gene* **51,** 247–254.
16. Messing, J. (1983) *Methods Enzymol.* **101,** 20–78.
17. Baralle, F. E. & Brownlee, G. G. (1978) *Nature (London)* **274,** 84–87.
18. Bader, B. L., Magin, T. M., Hatzfeld, M. & Franke, W. W. (1986) *EMBO J.* **5,** 1865–1875.
19. Chou, P. Y. & Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47,** 251–276.
20. Garnier, J., Osguthorpe, D. J. & Robson, B. (1978) *J. Mol. Biol.* **120,** 97–120.
21. Geisler, N. & Weber, K. (1982) *EMBO J.* **1,** 1649–1656.
22. Crick, F. H. C. (1953) *Acta Cryst.* **6,** 689–697.
23. Raychaudhury, A., Marchuk, D., Lindhurst, M. & Fuchs, E. (1986) *Mol. Cell. Biol.* **6,** 539–548.
24. Marchuk, D., McCrohon, S. & Fuchs, E. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1609–1613.
25. Steinert, P. M., Parry, D. A. D., Idler, W. W., Johnson, L. D., Steven, A. C. & Roop, D. R. (1985) *J. Biol. Chem.* **260,** 7142–7149.
26. Johnson, L. D., Idler, W. W., Zhou, X.-M., Roop, D. R. & Steinert, P. M. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1896–1900.
27. Jorcano, J. L., Rieger, M., Franz, J. K., Schiller, D. L., Moll, R. & Franke, W. W. (1984) *J. Mol. Biol.* **179,** 257–281.
28. Hatzfeld, M. & Franke, W. W. (1985) *J. Cell Biol.* **101,** 1826–1841.
29. Eichner, R., Sun, T.-T. & Aebi, U. (1986) *J. Cell Biol.* **102,** 1767–1777.