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 \approx 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 struc-

ture and a very high sequence identity at the DNA and amino acid sequence level to the corresponding bovine keratin.*

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 *EcoRI/Sma* I adaptors, and cloned into *EcoRI*-digested λ gt10 (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 [³²P]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 \approx 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-

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.

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).

H40 CCAJ GGCC	<u>40</u> CGCCGGACCGGGGCGGGGGCACCTCTGGAGGGCAGGGCCTCTGGAGGGAG															66 165 264									
<u>ат</u> д И	ACT T	тсс s	TAC Y	AGC S	TAT Y	CGC R	CAG Q	TCG S	TCG S	GCC À	ACG T	TCG S	тсс s	TTC F	GGA G	GGC G	CTG L	GGC G	GGC G	GGC G	тсс s	GTG V	CGT R	TTT F	339 25
GGG G	CCG P	GGG G	GTC V	GCT À	TTT F	CGC R	GCG À	CCC P	AGC S	ATT I	сас Н	GGG G	GGC G	тсс s	GGC G	GGC G	CGC R	GGC G	GTA V	тсс s	GTG V	TCC S	тсс s	GCC A	414 50
CGC R	TTT F	GTG V	тсс s	TCG S	тсс s	тсс s	TCG S	GGG G	GGC G	TAC Y	GGC G	GGC G	GGC G	TAC Y	GGC G	GGC G	GTC V	CTG L	ACC T	GCG A	тсс s	GAC <u>D</u>	GGG -G	CTG L-	489 75
CTG -L-	ссс <u>А</u>	GGC G-	AAC	GAG	AAG	CTA	ACC	АТG <u>М</u>	CAG	AAC	стс <u>L</u>	AAC	GAC	CGC	СТG <u>L</u>	GCC λ	тсс S	TAC <u>Y</u>	CTG	GAC	AAG	бтс <u>V</u>	CGC	GCC A -	564 100
CTG -L	GAG <u>E</u>	GCG	GCC A	аас <u>N</u>	GGC G	GAG	ста <u>L</u>	GAG	GTG V	AAG	атс <u>1</u>	CGC	GAC	тсс <u>₩</u>	TAC Y	CAG Q	AAG K	CAG Q	GGG	CCT P	GGG G	CCC P	TCC S	CGC R	639 125
GAC D	TAC Y	AGC S	САС Н	TAC Y-	TAC	ACG	ACC	АТС I-	CAG	GAC	CTG	CGG	GAC	AAG	ATT I-	CTT	GGT	GCC	ACC	ATT	GAG	AAC	TCC	AGG	714 150
атт - <u>1</u> -	GTC V-	CTG L-	С л G Q-	атс <u>1</u> -	GAC	AAT	ссс <u>А</u>	CGT	CTG	GСТ А	са <u>∆</u>	GAT	GAC	ттс <u>F</u> -	CGA	ACC	AAG	<u>F</u>	GAG	ACG	даа <u>Е</u> -	CAG	GCT	CTG L-	789 175
CGC - <u>R</u> -	АТG М -	AGC	бтс ⊻-	GAG	GCC	GAC	атс 1-	AAC	GGC	CTG	CGC	AGG	GTG V-	стс <u>L</u> -	GAT	GAG	CTG	ACC	CTG	GCC A -	agg <u>R</u> -	ACC	GAC	стс <u>L</u> -	864 200
GAG -E-	АТG M -	CAG Q-	атс <u>1</u> -	даа Е-	GGC	стс <u>L</u> -	AAG	GAA E-	GAG	стс <u>L</u>	GCC λ -	тас у-	стс <u>L</u> -	AAG K-	AA G K-	AAC	сат <u>н</u> -	GAG	GAG	даа <u>е</u> -	А ТС 1-	A GT S-	ACG	стс <u>L</u> -	939 225
AGG R	GGC G	CAA Q	GTG V	GGA G	GGC G	CAG Q SP	GTC V acer	AGT S 2	GTG V	GAG E	GTG V	GAT D	тсс s	GCT À	CCG P	GGC G	ACC T	GAT D	CTC L	ссс д-	ААG К-	АТС I-	стс <u>L</u> -	AGT S-	1014 250
GAC	ATG	CGA <u>R</u>	AGC	САА Q-	тат <u>¥</u>	GAG	GTC	ATG	ссс <u>А</u>	GAG	CAG	аа с <u>N</u>	CGG	АЛ С К-	дат D-	сст <u>à</u> -	GAA E-	GCC A	тсс W	ттс <u>г</u>	ACC	AGC	CGG R-	аст Т-	1089 275
GAA -E-	GAA	ттс <u>L</u>	AAC	CGG R-	GAG	бтс <u>V</u> -	сст а-	GGC G-	сас <u>н</u>	ACG	GAG	CAG	стс <u>L</u>	CAG	АТG М -	AGC <u>S</u> -	AGG	тсс s	GAG	GТТ ⊻-	аст т-	GAC	CTG L-	CGG R-	1164 300
CGC -R-	ACC	стт <u>L</u> -	САG Q-	GGT G-	стт <u>L</u> -	GAG E- co	ATT I- il 2	GAG	стс <u>L</u>	CAG	ТСА S	САG <u>Q</u>	CTG	AGC	АТG М-	ааа <u>к</u> -	сст л -	GCC	ттс <u>L</u>	GAA	GAC	аса т-	стс <u>L</u> -	GCA A-	1239 325
GAA -E-	ACG	GAG	GCG	CGC R-	ттт <u>F</u> -	GGA G-	GCC	С А G Q-	стс <u>L</u>	GCG	САТ н	атс <u>1</u>	CAG	GCG A-	CTG L-	атс <u>1</u> -	AGC	GGT G-	атт <u>1</u>	GAA E-	GCC A -	CAG Q-	стс <u>L</u> -	GGC G-	1314 350
GAT	стс ⊻-	CGA R-	сст л -	да т D-	аст <u>5</u> -	GAG	CGG	сад <u>Q</u> -	AAT	CAG Q	GAG	тас <u>1</u>	CAG	CGG R-	стс <u>L</u> -	АТG М -	GAC	АТС 1-	алс <u>К</u> -	TCG S-	CGG R-	стс <u>L</u> -	GAG	CAG Q-	1389 375
GAG -E-	атт 1-	GCC A -	АСС Т-	тас <u>1</u> -	CGC	AGC	CTG L-	стс <u>L</u> -	GAG	GGA G	САG <u>Q</u> =	GAA E- hel:	GAT D-	CAC ==H= ext	TAC == <u>¥</u> = ensi	AAC ==N= on	аат ==N=	ттс == <u>L</u> =:	TCT ==S=:	ссс == λ =	TCC ==S=	аас == <u>к</u> =	GTC ==V=	CTC ==L=	1464 400
TGA	GGC	AGCA	GGCT	CTGG	GGCT	тстс	CTGT	сстт	TGGA	GGGT	STCT	ICTG(GGTA	GAGG	GATG	GGAA	GGAA	GGGA	ссст	TACC	cccg	GCTC	TTCT	CCTG	1562

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

	1102030405060708090
H40 B40	Interesting
H46 H50	MTTCS <u>RQFTSSSS</u> MKGSCGIGGGIGAGSSRISSVLAGASCPASTYGGA®VSSRFSSGGACGLGGGYGGGFSSSS®FGSG <u>FGGGYGGGI</u> GA MTTCS <u>RQFTSSSS</u> MKGSCGIGGGIGAGSSRISSVLAGGSCRAPNTYGGGLSVSSSRFSSGGAYGLGGGYGGGFS®SSSSFGSGGGGGG
	coil 1A spacer 1
56 56 91 91	SSSGEYGGGYGGVLTASDGLLAGNEKLTMONINDRIASYLDKVRALEAANGELEVKIRDWYOKOGPGPSRDYSHYYTTIODIRDKILGAT SS.GGYGGGYGGALATSDGLLAGNEKLTMONINDRIASYLEKVRALEEANGDLEVKIRDWYOKOGPGPSRDYSHYYTTIODIRDKILGAT GFGGGLGAGFGGGFAGGDGLLVGSEKVTMONINDRIASYLDKVRALEEANADLEVKIRDWYORORPSEIKDYSPYFKTIEDIRNKIIAAT GLGAGLGGGFGGGFAGGDGLLVGSEKVTMONINDRIASYLDKVRALEEANADLEVKIRDWYORORPSEIKDYSPYFKTIEDIRNKIITAT
	coil 1B
146 145 181 181	IENSRIVLQIDNARLAADDPRTKPETEQALRNSVEADINGLRRVLDELTLARTDLENQIEGLKEELAVLKKNHEEEISTLRGOVGGOVSV IENSKIVLQIDNARLAADDPRTKPETEQALRNSVEADINGLRRVLDELTLARTDLENQIEGLKEELAVLKKNHEEEMSVLKGOVGGOVSV IENAHALLQIDNARLAADDPRTKYEARTGLRQTVEADVNGLRRVLDELTLARTDLENQIEGLKEELAVLKKNHEEEMLALRGOTGGDVNV VDNANVLLQIDNARLAADDPRTKYETELNLRNSVEADINGLRRVLDELTLARADLENQIESLKEELAVLKKNHEEEMNALRGOVGGOVSV
	spacer 2 coil 2
236 235 271 271	EVDSAPGTDLAKILSDHRSQVEVNAEQNRKDAEAWFTSRTEELNREVAGHTEQLQMSRSBVTDLRRTLQGLEIELQSQLSHKAALEDTLA EVDSAPGIDLAKILSDHRSQVEVIAEKNRKDAEAWFISQTSRLNREVAGHTEQLQISKTEVTDLRRTLQGLEIELQSQLSHKAALEGTLA EMDAAPGVDLSRILNENRDQVEQNAEKNRRDAETWFLSKTERLNREVASNSELVQSSRSBVTELRRVLQGLEIELQSQLSHKAALEGTLA EMDAAPGVDLSRILNENRDQVEQNAEKNRKDAEEWFFKTERLNREVATNSELVQSGKSEISELRRTMQNLEIELQSQLSHKAALENSLE
	coil 2 HE
326 325 361 361	ETEARPGAQLAHIQALISGIFAQLGDVRADSERONQEYQRINDIKSRLEQEIATVRSLLEGGEDHYNNLSASKVI.* ETEARPGAQLAQIQALISGIFAQLSDVRADTERONQEYQHINDIKTRLEQEIATVRNLLEGGDAYFNDLSLAKAI.* ETKGRYCMQLSQIGGLIGSVEGLAQLRCEMEQOSOBYQILLDVKTRLEQEIATVRRLLEGEDAHLSSQQASGQSYSSREVFTSSSSSSA ETKGRYCMQLAQIQEMIGSVEGLAQLRCEMEQONOBYKILLDVKTRLEQEIATVRRLLEGEDAHLSSSQFSGSQSSRDVTSSSRQIRT
451	VRPGPSSEQSSSSFSQGQSS*
451	KVMDVHDGKVVSTHEOVLRTKN*

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 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]	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).

1	1020																								
<u>H40</u> B40		CGCGGACCGGGGGGGGGGGCACCTCTGGAGGGCAGGGGCCTCTGGTCTCTGGAGGGGAGGGA															66								
CCAJ CCAJ	ANTGGGGAGAGAGCCCATATTTGCTCTCAGGAGCCTGCAAATTCCTCAGGGCTCAGATATCCGCCCCTGACACCATTCCTCCCCTCCCCCCCC															ACC GCC	165								
G GC(C GC(CCGCGGGG <u>CATAAAA</u> GGCGCCAGGTGAGGGCCTCGCCGCGCTCCCCGCGAATCGCAGCTTCTGAGACCAGGGTTGCTCCGTCCG															900 900	264								
ATG ATG	ACT ACT	TCC TCC	TAC TAC	XQC XQC	тат Тат	CGC CGC	сла Сла	TCG TCG	TCG TCG	G CC T CC	ACG ACC	TCQ TCQ	TCC TCC	TTC TTC	QQA QQG	GGC GGT	C TG A TG	99C 99C	99C 99C	99C 99C	TCC TCC	G TQ A TQ	CGT CGC	TTT TTC	339
900 900	C C G G C T	GGG GGA	GTC GGC	GCT GCC	TTT TTC	CGC CGC	oco oco	000 000	AGC AGC	ATT ATC	CĂC CĂT	900 900	ggc ggc	TCC TCA	GGC GGT	990 990	CGC CGC	99C 99C	GT A GT G	TCC TCG	GTQ GTQ	TCC TCC	TCC TCC	900 900	414
coc coc	TTT TTC	GTG GTG	TCC TCC	TCG TCG	TCC TCC	TCC TCC	TCG	999 999	GGC GGC	тас Тас	GGC GGC	GGC GGC	99C	TAC TAT	GGC GGG	GGC GGC	GTC GCC	CTG CTG	ACC GCC	G C G ACC	TCC TCC	сас Сас	999 999	CTQ CTQ	489
CTG CTG	aca aca	GGC GGC	аас Аас	gaq Gag	AAQ AAQ	CTA CTC	ACC ACC	ATG Atg	сло Сло	алс Алс	CTC CTC	алс Алс	сус Сус	coc coc	CTG CTG	GCC GCC	TCC TCC	тас Тас	CTG CTG	GAC GAG	алс Алс	oto oto	cec	900 900	564
CTQ CTQ	gag gag	GCG GAG	900 900	AAC AAC	99C 99C	GAG GAC	CTA TTG	gag Gag	GTG GTG	aag aag	ATC ATC	CGC CGC	дас Дас	TQQ TQQ	тас Тас	CAG CAG	aag aag	слс Слс	000 000	CCT CCC	000 000	000 000	TCC GCC	coc coc	639
өлс өлс	TAC TAC	AGC AGC	CAC CAC	TAC TAC	TAC TTC	ACQ AAQ	ACC ACC	ATC ATA	CAG GAG	GAC GAC	CTG CTG	cgg cgg	дас Дас	AAG Cag	ATT ATT	CTT CTC	got Got	900 900	ACC ACC	λTT λTT	gag gag	AAC AAC	TCC	AGQ Aaq	714
ATT ATA	GTC GTC	CTG CTG	сло Сло	ATC ATC	GAC GAC	777 772	acc acc	CGT CGT	CTG CTG	oct oct	GCX GCX	GAT Gat	GAC GAC	TTC TTC	CGA CGC	ACC ACC	алс Алс	111 111	gag Gag	ACG ACG	GAA GAG	CAG CAA	oct oct	C TG T TG	789
coc coc	ATG ATG	AGC AGT	GTQ GTQ	слс Слс	GCC GCT	дас Сас	ATC ATC	AAC AAT	99C	CTG CTG	cec cec	A gg C gg	GTG GTG	CTG CTG	GAT GAT	GAG GAG	CTG CTG	ACC ACC	CTG CTG	900 900	XQQ XQQ	ACC ACT	сус Сус	CTG CTG	864
GAG GAG	ATG ATG	сла Сла	ATC ATT	сал Сла	GGC GGC	CTG CTG	алс Алс	GAA GAG	слс Слс	CTG CTG	ecc ecc	ТАС ТАС	CTG CTG	AAG Aag	алс Алс	аас Аас	CAT CAC	GAG GAG	gag Gag	сал Сал	ATC ATG	AGT AGT	ACC GTC	CTG CTG	939
AGG Aag	ooc ooc	CAA CAG	oto Gtg	GGA GGT	GGC GGC	сло Сло	GTC GTC	AGT AGT	GTG GTG	сас Сас	GTG GTG	сат Сат	TCC TCT	GCT GCT	000 000	GGC GGC	ACC ATC	GAT GAC	CTC CTA	GCC GCC	алс Алс	ATC ATC	CTG CTG	agt agt	1014
GAC GAC	ATG ATG	C GA A GA	XGC XGC	сал Сал	TAT TAT	gag gag	GTC GTC	ATG ATT	GCC GCT	GAG GAG	САG Алс	AAC AAC	CGG CGG	алс Алс	GAT Gat	GCT GCT	GAA GAG	ecc ecc	TGG TGG	TTC TTC	ACC ATC	AGC AGC	CGQ Caq	ACT ACC	1089
GAA GAG	GAA GAG	TTG CTG	AAC	C QQ A QQ	GXG GXG	GTC GTC	GCT GCT	GGC GGC	САС САС	ACG ACG	сас Сас	CAG CAG	CTC CTG	сад Сад	ATG ATC	AGC AGC	AGQ AAQ	TCC ACG	слс Слс	GT T GT C	ACT ACC	бус Сус	CTG CTG	CGQ CGQ	1164
coc coc	ACC	CTT CTC	сло Сло	ggt ggt	CTT CTG	сус Сус	ATT	gag gag	CTG CTG	CAG CAG	TCA TCT	сло Сло	CTG CTC	AGC AGC	ATG Atg	ала Ала	GCT GCC	ecc ecc	TTG CTG	сал Сал	GAC GGC	аса Аса	CTQ CTQ	GCA GCG	1239
GAA GAA	ACQ	gag gag	GCG GCT	CGC	TTT TTC	gga gga	900 900	сло Сло	CTG CTG	ece ece	CAT CAG	ATC ATC	сло сло	909 909	CTQ CTQ	ATC ATC	AGC AGT	ogt ggt	ATT ATC	сал Сал	900 900	сло Сло	CTQ CTQ	GGC AGT	1314
GAT GAC	GTQ GTQ	CGA CGA	OCT OCT	GAT	AGT	gag gag	C00 C00	сло Сло		сло Сло	сло сло	TAC TAC	CAG CAA	CGG CAC	CTC CTC	ATG ATG	GAC GAC	ATC ATC	ало Ало	TCG ACC	caa	CTQ CTQ	gag gag	CAQ CAQ	1389
gag gag	ATT ATT	GCC GCC	ACC	TAC	CGC CGA	AGC	CTG CTG	CTC CTG	GAG GAG	GGA GGC	сло сло	GAA GAT	GAT GCC	CAC TAC	TAC TTC	AAC AAC	AAT GAC	TTC CTC	TCT TCA	GCC CTC	TCC GCG	ала Ала	OTC OCC	CTC CTC	1464
TGA TGA	GTC	TGGC	GGCA GGCC	GCAG	GCTC	TGGG	GCTT CGCT	CTGC CTGC	TGTC TG···	CTTT ···CA	GGAG Gaag	G GTG GAAT	TCTT TCTC	CTGG CTGG	GTAG GTAG	AGGG AGGC	ATGG ATGG	GAAG Gagg	gaag Ggag	GGAC GGAC	CCTT CCTT	ACCC	CCCG TTCC	CTCT	1555

ТСТССТВАССТВССААТАААААТТТАТВВТССААБВВААААААААА 1602 ССССТВАССТВССААТААААСТТТАТВВСССАААВВАААААА

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 $\approx 35\%$ 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.

- 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.
- 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.
- Moll, R., Franke, W. W., Schiller, D. L., Geiger, B. & Krepler, R. (1982) Cell 31, 11-24.
- 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.
- 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.
- 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.
- Bader, B. L., Magin, T. M., Hatzfeld, M. & Franke, W. W. (1986) EMBO J. 5, 1865–1875.
- Chou, P. Y. & Fasman, G. D. (1978) Annu. Rev. Biochem. 47, 251-276.
- 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.
- Raychaudhury, A., Marchuk, D., Lindhurst, M. & Fuchs, E. (1986) Mol. Cell. Biol. 6, 539-548.
- Marchuk, D., McCrohon, S. & Fuchs, E. (1985) Proc. Natl. Acad. Sci. USA 82, 1609–1613.
- 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.
- Johnson, L. D., Idler, W. W., Zhou, X.-M., Roop, D. R. & Steinert, P. M. (1985) Proc. Natl. Acad. Sci. USA 82, 1896– 1900.
- 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.
- Eichner, R., Sun, T.-T. & Aebi, U. (1986) J. Cell Biol. 102, 1767–1777.