Conservation of the sizes of 53 introns and over 100 intronic sequences for the binding of common transcription factors in the human and mouse genes for type II procollagen (COL2A1)

Leena ALA-KOKKO,*† Ari-Pekka KVIST,* Marjo METSÄRANTA,‡ Kari I. KIVIRIKKO,* Benoit DE CROMBRUGGHE,§ Darwin J. PROCKOP†|| and Eero VUORIO‡

*Collagen Research Unit, Biocenter and Department of Medical Biochemistry, University of Oulu, Oulu, Finland, †Department of Biochemistry and Molecular Biology, Jefferson Institute of Molecular Medicine, Jefferson Medical College, Thomas Jefferson University, Philadelphia, PA 19107, U.S.A.,
†Departments of Medical Biochemistry and Molecular Biology, University of Turku, Turku, Finland,
and §Department of Molecular Genetics, University of Texas M.D. Anderson Cancer Center, Houston, TX, U.S.A.

Over 11000 bp of previously undefined sequences of the human *COL2A1* gene were defined. The results made it possible to compare the intron structures of a highly complex gene from man and mouse. Surprisingly, the sizes of the 53 introns of the two genes were highly conserved with a mean difference of 13%. After alignment of the sequences, 69% of the intron sequences were identical. The introns contained consensus sequences for the binding of over 100 different transcription factors that were conserved in the introns of the two genes. The first intron of the gene contained 80 conserved consensus sequences and the remaining 52 introns of the gene contained 106 conserved sequences for the binding of transcription factors. The 5′-end of intron 2 in both genes had a potential for forming a stem loop in RNA transcripts.

INTRODUCTION

Since the discovery of intronic sequences in genes over 15 years ago, there has been considerable speculation about the origin and the function of introns. Two alternative hypotheses have been advanced. One is that introns were present in genes when they first evolved, and they were retained in eukaryotes but lost in organisms subject to selective pressure on the volume of their DNA. The alternative hypothesis is that introns were added to genes during evolution and, as a corollary, they have been useful in shuffling exons to generate new gene and protein structures. In spite of intensive study, there have been no definitive data to support either of these hypotheses. In addition, several observations suggest that some intron sequences may constitute important regulatory elements for gene expression. For example, enhancer or silencer functions have been observed in sequences within the first intron of several collagen genes [1–4], in the single intron present in genes coding for chicken feather keratins [5], in the first intron of the c-*myc* gene [6], and in the first intron of the human growth hormone gene [7].

One test of the importance of gene structures is their conservation among species. Such comparisons have demonstrated that the number and positions of introns have been conserved in most genes since the vertebrate radiation about 100 million years ago. The nucleotide sequences with introns have been compared only for a few mammalian genes. The results have indicated that conservation of intron nucleotide sequences between mouse and man ranges from about 40% in some α-globin and β-globin genes to as high as 70% among the human acidic (type I) cytokeratins [8]. Most of the comparisons to date, however, have focused on genes with only a few introns. In this context, the introns of genes coding for the major fibrillar collagens provide an attractive target for comparison of intron sequences, since each gene contains 51–54 exons and the positions of the exons within each gene are highly conserved [9–11]. In addition, there is conservation of an unusual gene structure in that the large triple-helical domain of the α-chain of each collagen is coded for by 41 or 42 exons that are 54 bp, 45 bp or multiples of 54 bp and 45 bp; each exon begins with a complete codon for a glycine residue that appears as every third amino acid in the Gly-Xaa-Yaa- repeating structure of the α-chains; and there is conservation in the size and position of the four exons coding for the C-terminal propeptides that are found on the precursor pro-α chains of the proteins and that direct chain assembly. The only apparent variation in the gene structures appears in the variable number and length of the five to seven exons coding for the N-terminal propeptides of the proteins.

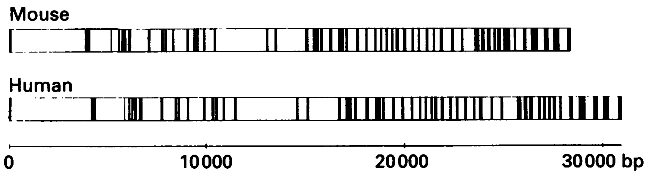

Figure 1 Exon–intron organization of the mouse and human type II procollagen genes

Black areas denote exons and white areas denote introns, drawn to the scale shown.

|| To whom correspondence should be addressed.

The nucleotide sequence data reported have been deposited in the GenBank/EMBL/DDBJ Nucleotide Sequence Databases under accession no. L10347.

Table 1 Exon–intron sizes of human, mouse and chick type II procollagen genes, and pairwise comparison of the human and mouse intron sequences

Mo, mouse; Hu, Human; Ch, chick. The exons and introns are numbered from 5 -end as in [10]. The sizes of exons and introns (in bp) are from references listed in the Materials and Methods section. Sizes of exons 1 and 52 refer to coding potential. The results of the overall homology analyses are given as overall percentage sequence identity .(%), the quality of alignments taking into account the length of the sequence, the number and length of the gaps (Q), and the mean ± SD of 100 randomized alignments as an estimate of the probability of obtaining the same sequence by chance (R). All the differences between Q and R are statistically significant.

	Exon size		Intron size					
No	Mo	Hu	Mo	Hu	Ch	%	Q	R ± SD
1	85	85	3799	4105		75.4	2358.4	1572 ± 9
2	204	207	1110	1494		65.9	601.5	465.6 ± 5.3
3	17	17	387	213		64.8	114.9	94.7 ± 2.4
4A	33	33	120	104		65.1	56.6	42.8 ± 2.1
4B	33	33	100	105		70.0	64.5	42.2 ± 2.2
5A	54	54	141	163		70.3	80.6	58.8 ± 2.6
5B	105	102	909	978		76.4	635.9	373.7 ± 4.9
6	78	78	619	627		75.7	440.1	250.6 ± 4.2
7	45	45	105	111	100	69.6	60.4	43.6 ± 2.4
8	54	54	336	400	124	73.4	188.3	142.9 ± 3.3
9	54	54	687	775	106	71.0	404.6	292.1 ± 4.8
10	54	54	290	374	101	66.0	137.8	123.3 ± 3.3
11	54	54	94	131	98	62.0	50.6	40.6 ± 2.4
12	54	54	291	306	85	71.2	167.1	117.2 ± 3.4
13	45	45	473	535	111	69.2	227.5	197.2 ± 3.8
14	54	54	2613	3062	90	62.6	1221.3	1098 ± 7.3
15	45	45	379	479	118	70.4	190.7	159 ± 3.3
16	54	54	1495	1513	78	69.4	864.2	607.7 ± 6.3
17	99	99	291	296	200	70.0	170.9	117.5 ± 3.9
18	45	45	87	92	374	72.4	58.5	37.6 ± 2.5
19	99	99	145	189	890	71.8	80.9	61 ± 2.7
20	54	54	371	524	411	64.6	193.5	164 ± 3
21	108	108	360	365	88	66.8	182.2	146.4 ± 3.9
22	54	54	84	84	89	69.5	55.4	33.9 ± 2.1
23	99	99	138	138	644	75.0	70.2	57 ± 2.7
24	54	54	465	440	82	68.4	247.1	181.1 ± 3.9
25	99	99	379	393	82	65.5	210.2	153.7 ± 4.1
26	54	54	392	406	115	60.3	199.2	158.7 ± 4.2
27	54	54	246	348	312	70.9	129.5	108.5 ± 2.8
28	54	54	228	243	78	72.4	149.7	96.9 ± 2.7
29	54	54	233	242	164	72.4	131.7	94.8 ± 3
30	45	45	186	146	270	73.1	82.4	68.1 ± 2.9
31	99	99	251	234	92	70.5	137.7	100.3 ± 3.3
32	108	108	338	342	80	59.7	168.0	134.4 ± 3.7
33	54	54	255	278	81	68.9	127.0	105.5 ± 3.6
34	54	54	361	376	81	63.7	176.8	144.8 ± 4.3
35	54	54	281	372	94	66.0	136.7	123.9 ± 3.3
36	54	54	283	254	89	63.1	131.3	106.2 ± 2.7
37	108	108	471	491	108	64.1	245.1	194.9 ± 4.4
38	54	54	435	444	114	61.8	214.1	175.8 ± 3.2
39	54	54	627	746	265	68.3	305.7	265.9 ± 4.6
40	162	162	169	197	91	66.5	94.8	73 ± 3
41	108	108	194	170	284	71.6	100.2	73.4 ± 2.6
42	108	108	217	354	158	68.4	125.7	100.1 ± 2.6
43	54	54	135	172	97	74.8	81.9	58 ± 2.3
44	108	108	190	165	274	65.6	84.3	69.4 ± 2.9
45	54	54	112	181	90	66.0	57.0	46.3 ± 1.7
46	108	108	234	244	140	67.9	123.4	94 ± 3.4
47	54	54	432	443	192	67.0	227.3	174.3 ± 4
48	108	108	285	357	78	73.1	168.2	121.5 ± 3.1
49	289	289	352	454	397	70.1	184.0	152 ± 3
50	188	188	297	343	593	72.9	182.0	124 ± 3.3
51	243	243	456	535	112	68.9	271.7	195.2 ± 4.5
52	147	147						

Recently we have completed the nucleotide structure of the 30 kb genes for type II procollagen (COL2A1) from both man and mouse. The data enable us to examine the degree of conservation of the nucleotide sequences of the introns during evolution of these two species.

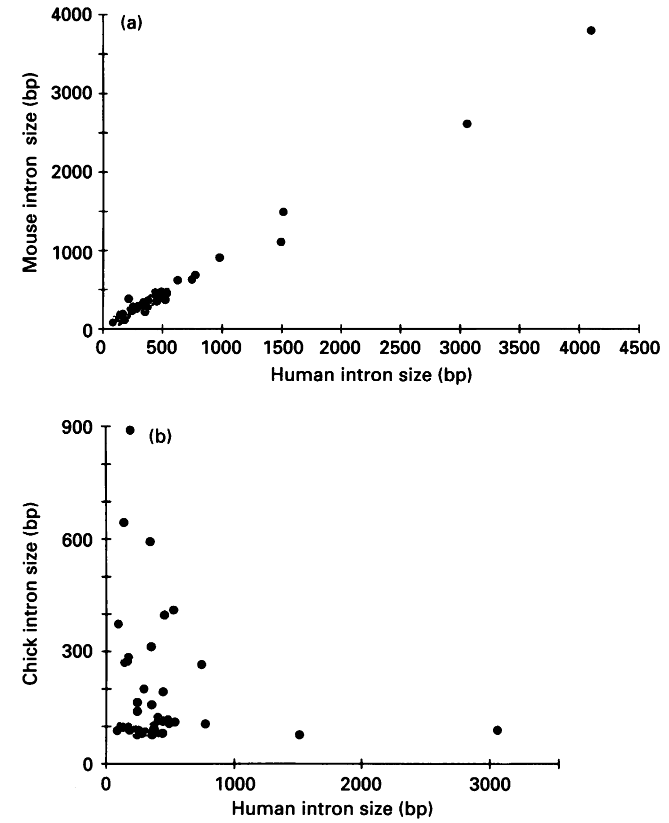

Figure 2 Graphical presentation of the pairwise size comparison of intron sizes between mouse and human (a) and chick and human (b) type II procollagen genes

MATERIALS AND METHODS

Nucleotide sequences

The nucleotide sequences for the complete mouse *COL2A1* gene were from Metsäranta et al. [12]. The accession number for the sequences is GenBank/EMBL M65161. Previously published nucleotide sequences for the human *COL2A1* gene were assembled from several sources, listed from the 5′-end to 3′-end: Vikkula et al. [13] (accession no. X58709); Huang et al. [14] (X57010; X57011); Ryan et al. [15]; Vikkula and Peltonen [16] (X16158); Cheah et al. [17]. The cDNA sequences for the human *COL2A1* gene were from Baldwin et al. [18] (X16711); and Elima et al. [19] (X06268; X02664; X06584).

The nucleotide sequencing of the human *COL2A1* gene in previously unsequenced areas was performed using genomic clones obtained with the cosmid cloning technique developed by us [20] followed by subcloning as described earlier [21]. The nucleotide sequencing was performed employing the dideoxynucleotide method and universal primers and specific oligonucleotides. The entire human *COL2A1* gene sequence from transcription start site to the translation stop codon has been assembled and deposited in the GenBank/EMBL database under the accession number L10347.

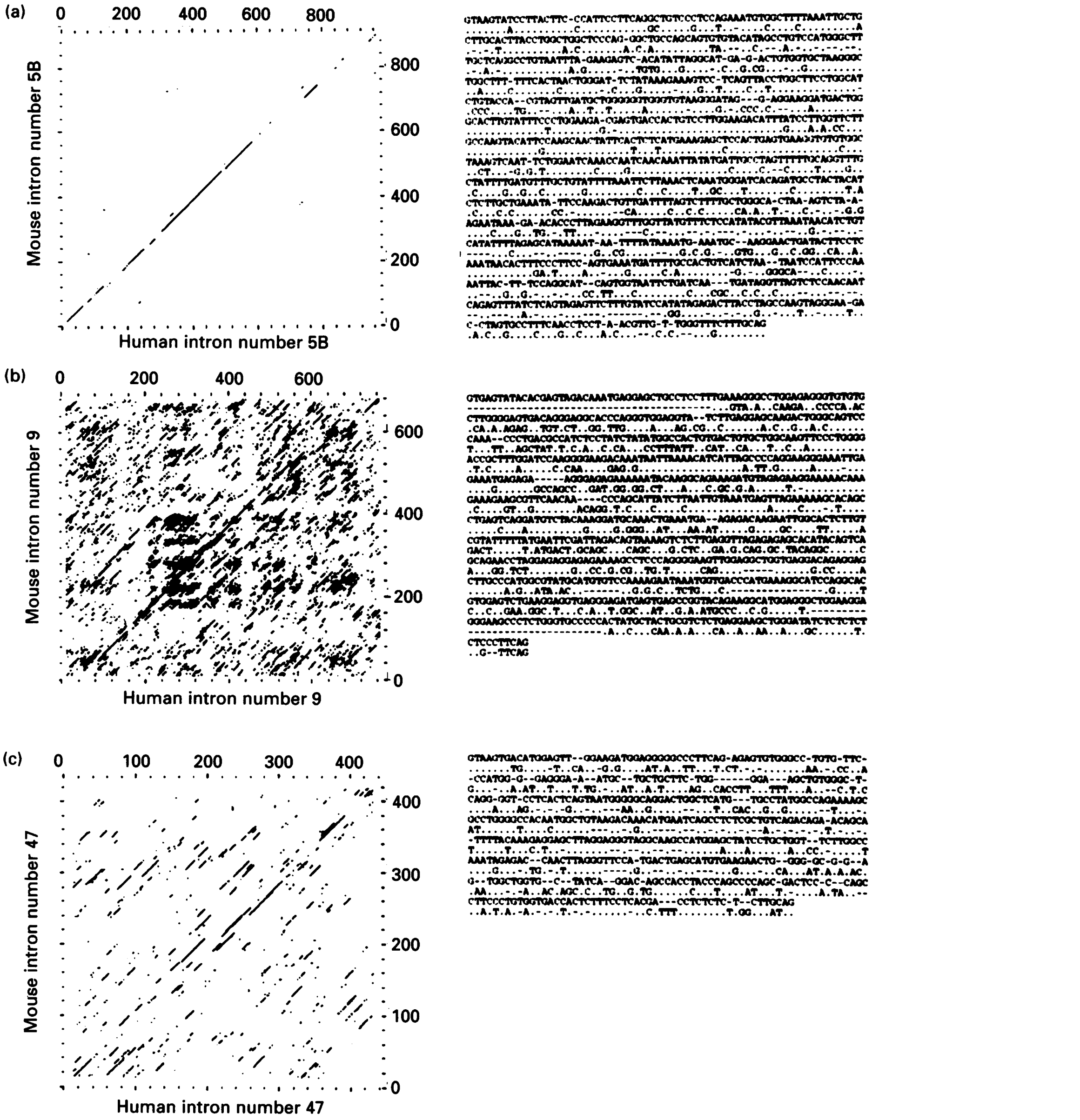

Figure 3 Dot matrix analyses and detailed comparison of maximally aligned intron sequences of introns 5B, 9, and 47

Top line, human sequence; bottom line, mouse sequence. Dashes indicate gaps introduced for maximal alignment. A dot indicates nucleotide identity.

Computer analysis of the nucleotide sequences

The GAP program from the University of Wisconsin GCG package (version 7.0; [22]) was used for overall comparisons. Consensus recognition sequences for transcription factors were searched with FINDPATTERNS. The dot matrix plot was created with COMPARE and DOTPLOT. Local identity of intron sequences was studied with MACAW program [23].

RESULTS

Completion of the nucleotide sequences of the human *COL2A1* gene

The complete nucleotide sequence of the *COL2A1* gene from mouse was reported earlier [12]. The complete cDNA sequence of the human *COL2A1* gene was also published earlier [18,19]. Also, 40 bp of sequences flanking each of 52 exons of the human *COL2A1* gene were published earlier [20]. Here over 11000 bp of additional intron sequences were defined so as to complete the 30997 bp structure of the human *COL2A1* gene. When discrepancies in the completed sequences were encountered, we used our own data, which in most cases came from sequencing of several alleles.

Comparison of exon and intron sizes

Analysis of the complete nucleotide sequences of the human and mouse *COL2A1* genes confirmed previous observations (for reviews, see [10,11]), that the exons were the same size except for small variations in the alternatively spliced exon 2 and the unusual exon 5B that is not found in genes for other fibrillar collagens (Figure 1 and Table 1). Exon 2 was 3 bp longer in the

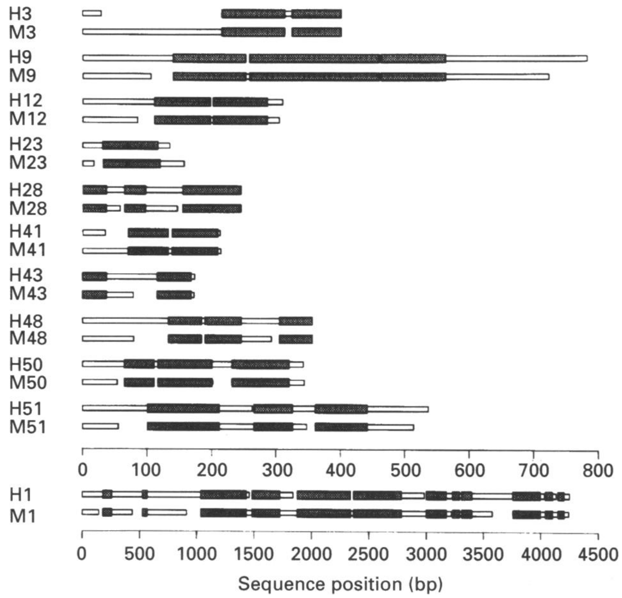

Figure 4 Representative examples of local identity alignments of intron sequences studied with the MACAW program

Intron numbers for human (H) and mouse (M) introns are shown on the left, and the scale in base pairs below. A different scale is used for intron 1 due to its large size. The shaded areas denote regions of significant sequence identity.

human gene and exon 5B was 3 bp longer in the mouse gene. Comparison of the intron size indicated that, although they were not identical in size, there was a higher degree of conservation of the size (Figure 2a). The mean difference in intron sizes was 12.9±1.5% (mean±S.E.M.). In contrast, there was little conservation of intron sizes when the human *COL2A1* gene was compared (Figure 2b) with the chick *COL2A1* gene [24].

Comparison of intron sequences

The GAP program of the GCG was used to compare the intron sequences of the mouse and human *COL2A1* genes. The MACAW program was used to align intron sequences.

After alignment of the sequences, the identity ranged from 60% to 75%, with a mean value of 69% (Table 1).

Three illustrative introns are presented by dot matrix analysis in Figure 3. Figure 3(a) presents intron 5B, an example of an intron with a high degree of sequence identity. Figure 3(b) presents intron 9, an example of an intron with a moderate degree of identity. Figure 3(c) presents intron 47 as an example of an intron with a relatively low degree of identity. Eleven introns with a high degree of local identity are illustrated in Figure 4.

Binding sequences for transcription factors

The intron sequences of the mouse and human *COL2A1* genes were also searched for consensus sequences for the binding of transcription factors. As indicated in Tables 2 and 3, consensus sequences conserved between the two species were found for the binding of over 100 different transcription factors. Many of the transcription factors had more than one conserved sequence for binding. The heaviest concentration of conserved sequences was in intron 1, which contained a total of 80 such conserved sequences (Table 2). The remaining 52 introns of the gene contained conserved consensus sequences for the binding of 106 transcription factors (Table 3). Multiple copies of several of the consensus sequences were present in both genes. For example, 23 conserved sequences in the human gene were found for the transcription factor malT-CS/Rev. Nineteen of these sequences were scattered in the mouse gene.

Stem loop formation by intron sequences

The complete nucleotide sequences of both the human and the mouse *COL2A1* genes were analysed for the potential of forming double-stranded stem loop structures in RNA transcripts of the gene. The FASTA and STEMLOOP/DOTLOT program [22] was used for the analysis. As the first step in the analysis, the reverse sequence of each gene was matched with the 5′ to 3′ sequence to find the longest base paired region. With both genes, the longest region was at the 5′-end of intron 2. As a second step, the 5′-ends of intron 2 were analysed for stem loop formation. As indicated in Figure 5, the sequences from the 5′-end of the intron of both genes formed stem loops. The stem loop from the human sequence began with the last two nucleotides of exon 2 and included the first 55 nucleotides of the intron. The stem loop of the mouse gene began with the first nucleotide of the intron and included 38 nucleotides. The ΔG value for the human stem loop was -54.0 kJ (-12.9 kcal) and the ΔG value for the mouse stem loop was -32.2 kJ (-7.7 kcal).

DISCUSSION

The results presented here provide a comparison of the 53 introns of the *COL2A1* gene between mouse and man. The data demonstrate that the intron sizes are well conserved between the two genes from mouse and man, but not between the *COL2A1* gene in these species and the same gene from chick. The nucleotide sequences of the introns of the *COL2A1* gene are also well conserved between mouse and man. The mean value of 69% identity suggests that the conservation of intron sequences ranks at the upper range of intron sequences from the two species that have been compared to date. This suggests some function for the conservation of sequences within introns. Although the chick type II collagen gene shares the same evolutionary and fractional relationship, very little sequence conservation was observable within introns. Yet the *COL2A1* gene exhibits very similar tissue-specific expression and undergoes similar processing in all three species. Interestingly, the sequence similarity decreases markedly about 200 bp beyond the polyadenylation site, suggesting that sequence conservation is limited to the transcription unit.

The large amount of sequence data available for comparison made it possible to carry out an extensive search for consensus sequences for the binding of known transcription factors. The results provided the surprising finding that the introns of the gene contained consensus sequences for binding of over 100 different transcription factors that were conserved in the two species. Conserved consensus sequences for a total of 80 known transcription factors were found in the first intron. This observation is consistent with several reports indicating that the first intron of the *COL2A1* gene, as well as the first intron of several other collagen genes, contains important regulatory sequences [1–4]. However, an additional 106 conserved sequences for the binding of transcription factors were found in the other 52 introns of the genes. The data obviously do not establish that these sequences are functionally important, but the results raise the possibility that these additional conserved sequences may be important in regulation of expression of the gene. This suggestion is consistent with the observations that elements in or near the 3′-

Table 2 Binding sites for transcription factors found in intron 1

Factor	Recognition sequence	Copies (Human/Mouse)	Factor	Recognition sequence	Copies (Human/Mouse)
(Sp1)-TK.1	CCCCGCCC	1/1	histone_H4_CS.2/Rev	TCAGGR	4/1
AABS_CS2	GTGNNGYAA	1/1	hsp70.2[h]	GGCGGG	4/3
Ad-conserved-sequence-e[a]	TGACGT	1/1	hsp70.2/Rev[i]	CCCGCC	6/5
aD-globin-CS2/Rev	CCGCACGG	1/1	IE1.2	CTTTCC	5/6
AP-2_CS	GSSWGSCC	3/3	IE1.2/Rev	GGAAAG	3/2
AP-2_CS/Rev	GGSCWSSC	2/3	INF.1/Rev	TCACTT	1/1
AP1_CS3/Rev	TKANTCA	1/3	insulin/Rev	TTTCCAC	3/1
AP2_CS4	YCSCCMNSSS	5/9	JCV_repeated_sequence	GGGNGGRR	1/5
AP2_CS4/Rev	SSSNKGGSGR	2/8	JCV_repeated_sequence/Rev	YYCCNCCC	10/15
bA-globin.1/Rev[b]	CCCGCCCC	1/1	lambda.c/Rev	CRYACRCC	1/1
BGP1_RS/Rev[c]	CCGCCC	4/4	malT-malPp	TCCTCC	6/4
CAP/CRP-lac/Rev	AAAGTGT	2/1	malT-malPp/Rev	GGAGGA	2/1
CK-8-mer/Rev	TTTGGNTT	1/1	malT_CS	GGAKGA	2/2
CTF/CBP-hs[d]	GATTGG	1/3	malT_CS/Rev	TCMTCC	7/5
E-alpha_H_box/Rev	CAGGTCC	1/1	MRE_CS2/Rev	GNGYGCA	1/1
E1A-F_CS	XGGAYGT	1/4	NF-E1_CS1	MYWATCWY	1/1
EARLY-SEQ1	YYCCGCCC	1/1	NF-E1_CS1/Rev	RWGATWRK	2/1
EBP20_CS1	TKNNGYAAK	4/3	NFI.2/Rev	GATTGGC	1/1
EBP20_CS1/Rev	MTTRCNNMA	4/1	NFkB_CS1/Rev	GDRRADYCCC	1/1
EBP20_CS3	TCNTACTC	1/1	NRE_Box1_CS	ANCCTCTCY	1/1
engrailed_CS/Rev	TTTDATWGD	2/1	NRE_Box1_CS/Rev	RGAGAGGNT	1/1
enhancer_core/Rev	CWWWCCAC	2/1	PEA3_RS/Rev	CTTCCT	2/1
ETFA.2/Rev[e]	TGACGTRR	1/1	PEBP2_RS/Rev	GCGGTC	1/1
FSE2.1/Rev	TCCTCTC	1/1	PRL_conserved_motif/Rev	TWWTCAGG	1/1
GAGA-en/Rev	CTCTCTCT	1/1	Pu_box/Rev	TTCCTC	3/3
GAL1-TATA/Rev	TTATAT	2/1	SDR_RS/Rev	CACCSCYC	2/1
GCN4-HIS/Rev[f]	GAGTCA	1/2	Sp1?-U2snR.2	ACGCCC	1/1
GCN4-HIS4.3	CAGTCA	1/2	SV40.13[j]	TGGAAAG	1/1
GCN4-HIS4.3/Rev	TGACTG	2/1	SV40.13/Rev[k]	CTTTCCA	2/3
GCN4-ILV1.3/Rev	TGACTT	1/1	TATA-box.1/Rev	ATTATA	1/1
GCN4-ILV2	TGATTC	2/1	TFIID-EIIa/Rev	TTTGTA	3/3
GCN4-ILV2/Rev	GAATCA	1/1	UBP1_RS	CTCTCTGG	1/1
GH-CSE2/Rev	ATTTATT	1/1	uteroglobin_HS-2.4_CS	RYYWSGTG	2/2
GR-MT-IIA/Rev	AGGACA	2/1	uteroglobin_HS-2.4_CS/Rev	CACSWRRY	2/1
H2A_conserved_US	YCATTC	2/3	WAP_US5	CCAAGT	1/1
H2B-CCAAT/Rev	TNATTGG	1/1	WAP_US6	TTTAAA	3/1
H4TF-1hist/Rev	GAAATC	2/2	XlHbox1-F1/Rev	TTTAATTG	1/1
HC3	CCACCA	2/3	XRE_CS1/Rev	WGCGTG	5/1
HC3/Rev	TGGTGG	6/1	yeast-termination-CS1/Rev	ACTANNNNNNNCTA	1/1
Hepta/Rev[g]	TCATGAG	1/1	zeste-Ubx/Rev	CGCTCG	1/1

[a] Ad-conserved-sequence-e = ATF_RS, CREB_CS/Rev
[b] bA-globin.1/Rev = Sp1-SV40.1/Rev
[c] BGP1_RS/Rev = LSF-SV40, Sp1-IE-3.1, Sp1-IE-3.2/Rev, Sp1-IE-3.4, Sp1-IE-3.5, Sp1-IE-4/5.2, Sp1-SV40.4/Rev, Sp1?-U2snR.1/Rev, SP1?-2snR.3/Rev, Sp1_CS2/Rev, SP1_CS3/Rev
[d] CTF/CPB-hs = hsp70.5
[e] ETFA.2/Rev = ETFA.3
[f] GCN4-HIS/Rev = GCN4-HIS3.2, GCN4-HIS3.5/Rev, GCN4-HIS4.1/Rev, GCN4-HIS4.2/Rev, GCN4-ILV1.1, GCRE/Rev
[g] Hepta/Rev = IgH-heptamer/Rev
[h] hsp70.2 = Sp1-hsp70, Sp1-IE-3.3/Rev, Sp1-IE-4/5/Rev
[i] hsp70.2/Rev = Sp1-hsp70/Rev. Sp1-IE-3.3, Sp1-IE-4/5
[j] SV40.13 = SV40.16, SV40.6
[k] SV40.13/Rev = SV40.16/Rev, SV40.6/Rev

end of the genes are important for tissue-specific expression of the β-actin gene [25], the *COL1A1* gene [26,27], the erythropoetin gene [28], the γ-globin gene [29,30], the human tumour necrosis factor gene [31] and the keratin 1 gene [32].

The observation that the 5′-sequences of intron 2 of both the mouse and human *COL2A1* gene have a potential for forming a stem loop is of interest, because exon 2 of the *COL2A1* gene is the only exon in a gene for a fibrillar collagen known to undergo alternative splicing [33]. RNA secondary structure is frequently cited as an explanation for alternative splicing of RNA transcripts but the suggestion has been difficult to prove [34–37]. The potential stem loop structures identified here were relatively long in that they involved 57 nucleotides in the human gene and 38 nucleotides in the mouse gene. Also, they were relatively stable with ΔG values of -54.0 and -32.2 kJ. It will probably be necessary, however, to develop further evidence for their potential role in the alternative splicing of exon 2.

The evolutionary relationship of all fibrillar collagen genes is clearly demonstrated by the highly conserved exon organization. While the location of introns has been maintained among species, the sizes of introns between different collagen genes show little conservation with the exception of intron 1, which is large in all the genes for major fibrillar collagens. The maintenance of intron sizes and sequence information observed here for the human and mouse *COL2A1* genes suggests that additional information may have been vested into the introns. Some of this information may

Table 3 Binding sites for transcription factors found in introns 2–52

Factor	Recognition sequence	Intron location (Human/Mouse)
Ad2MLP_US.3[a]	TATAAA	6(1/1)
Ad2MLP_US.3/Rev[b]	TTTATA	5B(1/1)
Adh1_US2/Rev	CCGGGG	14(1/1)
AP-2_CS	GSSWGSCC	14(6/2),16(1/2)
AP-2_CS/Rev	GGSCWSSC	14(4/4),16(3/1),30(2/1)
AP1_CS3	TGANTMA	2(1/3),14(1/3)
AP1_CS3/Rev	TKANTCA	14(1/2)
AP2_CS4	YCSCCMNSSS	16(1/1),20(3/1)
AP2_CS4/Rev	SSSNKGGSGR	39(1/1),43(1/1)
B-factor-hsp70	TATAAATA	6(1/1)
bA-globin.4/Rev	CCCCTCCTC	14(1/1)
c-mos_DS3/Rev	TTAAAAC	9(1/1)
CK-8-mer/Rev	TTTGGNTT	6(1/1)
CTF/CBP-hs/Rev[c]	CCAATC	14(2/1)
CuE3.1/Rev[d]	GCCACATG	49(1/1)
E1A-F_CS	XGGAYGT	9(1/1),14(1/2)
E1A-F_CS/Rev	ACRTCCX	16(1/2),36(1/1)
EBP20-ATsC2/Rev	GCTTAAGA	14(1/2)
EBP20_CS1	TKNNGYAAK	16(1/1),51(1/1)
EBP20_CS1/Rev	MTTRCNNMA	2(1/1),14(2/1),16(1/1), 50(1/1)
EFII-RSV/Rev	TGCATA	16(1/1)
element_II_rs-3	TTTGGCC	17(1/1)
element_II_rs-3/Rev	GGCCAAA	14(1/1)
FSE2.1	GAGAGGA	9(1/1),27(1/1)
GAGA-en	AGAGAGAG	9(1/1)
GAGA-en/Rev	CTCTCTCT	47(2/1)
GAGA_RS/Rev	GCTCTCTCK	4A(1/1)
GAL1-TATA/Rev	TTATAT	5B(1/1)
GCN4-HIS[e]	TGACTC	35(1/1)
GCN4-HIS/Rev[f]	GAGTCA	2(1/1),14(1/1)
GCN4-HIS3.4/Rev	GAGTAA	14(1/1)
GCN4-HIS4.3/Rev	TGACTG	5B(1/2),9(1/1),38(1/1)
GCN4-ILV1.2[g]	TGAGTG	2(6/1),5B(1/1),22(1/1), 35(1/1),40(1/1),41(1/1), 51(2/1)
GCN4-ILV1.2/Rev[h]	CACTCA	14(2/1)
GCN4-ILV1.3/Rev	TGACTT	21(1/1),39(1/1)
GCN4-ILV2	TGATTC	14(1/1)
GCN4-ILV2/Rev	GAATCA	5B(1/1)
GH-CSE	TAAATTA	2(1/1)
GH-CSE/Rev	TAATTTA	2(1/2)
GH-CSE2	AATAAAT	2(1/1)
GH-CSE2/Rev	ATTTATT	2(2/3)
GH1/Rev	CAGACAGA	37(1/1)

GR-MT-IIA	TGTCCT	5B(1/1),13(1/1),14(2/1)
		24(1/1),32(1/1),44(1/1),
		48(1/3)

MRE_CS2	TGCRCNC	14(1/1),28(1/1),37(1/1)
MRE_CS2/Rev	GNGYGCA	14(2/1),38(1/1)
MT-I.1	GCACTC	14(1/1),15(1/1)
MTVGRE_NRS	AGGATGT	9(1/1)
NF-E1.6	TATCTC	36(1/1)
NF-E1.8	CCAATCT	14(1/1)
NF-E1_CS1	MYWATCWY	2(1/1),8(1/1)
NF-E1_CS1/Rev	RWGATWRK	14(3/3)
NFI_CS3/Rev	TGGCNNNNNNCCA	5B(1/2),6(1/1)
NFkB_CS1	GGGRHTYYHC	33(1/1)
NFkB_CS4	GGGRNTYYC	16(1/3)
NRE_Box1_CS/Rev	RGAGAGGNT	9(1/1)
oli1_DS	ATTCTTA	5B(1/1)
PEA3_RS	AGGAAG	5B(1/2),6(1/1),9(2/1), 13(1/2),14(4/4),16(1/3), 26(1/1),27(1/1)
PEA3_RS/Rev	CTTCCT	2(3/4),5B(2/1),8(2/1), 14(5/4),20(1/2),34(1/2), 36(1/1),37(1/1)
PEBP2_RS	GACCGC	14(1/1)
GR-MT-IIA/Rev	AGGACA	42(2/1)
GR-uteroglobin.1[i]	TGTTCT	6(1/1),14(1/1)
GR-uteroglobin.1/Rev[j]	AGAACA	5B(1/1),12(1/1)
GRE_CS7[k]	WCTGWTCT	6(1/1)
H2A_conserved_US	YCATTC	5B(2/1),14(1/2),51(2/1)
H2A_conserved_US/Rev	GAATGR	14(1/1),24(2/1)
H2B-CCAAT/Rev	TNATTGG	26(1/1)
H4TF-1hist	GATTTC	14(1/1)
HC3	CCACCA	8(1/1),14(4/1),16(1/2), 49(1/1)
HC3/Rev	TGGTGG	2(2/1),14(1/3),21(1/2), 26(2/3)
HiNF-Ahist	AGAAATG	9(1/1)
histone_H4_CS.2	YCCTGA	9(1/1),13(1/1),14(4/4), 17(1/3),20(1/1),25(2/1), 32(1/1),38(1/1),39(2/1), 48(2/1)
histone_H4_CS.2/Rev	TCAGGR	9(2/1),14(3/3),16(3/1), 27(1/1),37(1/1),47(2/2)
hsp70.2[l]	GGCGGG	14(1/2)
HSV_IE_repeat	GCGGAA	16(1/1)
IE1.2	CTTTCC	2(2/2),5B(2/1),8(1/1), 10(2/1),14(2/1),25(1/1)
INF.1	AAGTGA	31(1/1)
INF.1/Rev	TCACTT	2(1/1),30(1/1)
JCV_repeated_sequence	GGGNGGRR	2(2/1),16(2/5),19(1/1), 26(1/1),31(2/3),34(1/1), 35(2/3),39(2/2),46(1/5), 47(2/1),51(1/1)
JCV_repeated_sequence/Rev	YYCCNCCC	8(1/1),14(3/6),16(1/4), 20(2/1),23(1/2),25(1/1), 48(2/2)
lambda.c	GGYGTRYG	5B(1/1),13(1/1)
lambda.d	GGTGTGTG	13(1/1)
LVa-Mo-MuLV	GAACAG	16(2/1)
LVa-Mo-MuLV/Rev[m]	CTGTTC	6(2/2),14(1/1),16(1/1)
LVa_RS	GAACAG	16(2/1)
LVb-Mo-MuLV/Rev	TATCCTG	14(1/1),44(1/1)
LVb_RS/Rev	TATCCTG	14(1/1),44(1/1)
malT-malPp	TCCTCC	2(1/1),13(1/1),14(2/2), 50(1/2)
malT-malPp/Rev	GGAGGA	16(1/3),26(1/1),31(2/1), 35(1/1),37(1/2),39(3/2)
malT_CS	GGAKGA	2(1/2),16(1/3),26(1/2), 31(2/1),35(1/1),37(1/2), 39(4/5),42(1/1)
malT_CS/Rev	TCMTCC	2(1/2),13(1/1),14(2/2), 15(3/1),16(1/1),20(1/1), 21(3/2),34(2/1),39(1/1), 50(1/2)
Pu_box	GAGGAA	6(1/1),9(1/1),14(3/1), 31(2/1)
Pu_box/Rev	TTCCTC	2(3/1),14(2/1),47(1/1)
retroviral_TATA/Rev	CTWWATWS	9(1/1)
rRNA-T2/T3[n]	GACTTGC	14(1/1)
STE2.1/Rev	AAGTACAT	5B(1/1)
SV40.13/Rev[o]	CTTTCCA	5B(1/1),14(1/1)
TATA-box-CS	TATAWAW	6(1/1)
TATA-box.1/Rev	ATTATA	5B(1/1)
TFIID-EIIa/Rev	TTTGTA	5B(1/1),8(1/1),10(1/1)
uteroglobin_HS-2.4_CS	RYYWSGTG	14(2/2),20(1/1),38(1/1), 42(1/2)
uteroglobin_HS-2.4_CS/Rev	CACSWRRY	32(1/1)
vaccinia-term-sequence	CTATTC	5B(1/1)
WAP_US5	CCAAGT	5B(2/1),14(2/1)
WAP_US5/Rev	ACTTGG	14(3/1)
WAP_US6	TTTAAA	5B(2/2),13(1/1)
XRE_CS1	CACGCW	2(1/2)

[a] Ad2MLP_US.3 = his3-Tr-TATA, TATA-box.2, TFIID/TBF?-RS

[b] Ad2MLP_US.3/Rev = his3-Tr-TATA/Rev, TATA-box.2/Rev, TFIID/TBF?-RS/Rev

[c] CTF/CBP-hs/Rev = hsp70.5/Rev

[d] CuE3.1/Rev = IgHC.8/Rev

[e] GCN4-HIS = GCN4-HIS3.2/Rev, GCN4-HIS3.5, GCN4-HIS4.1, GCN4-HIS4.2, GCN4-ILV1.1/Rev, GCRE

[f] GCN4-HIS/Rev = GCN4-HIS3.2, GCN4-HIS3.5/Rev, GCN4-HIS4.1/Rev, GCN4-HIS4.2/Rev, GCN4-ILV1.1, GCRE/Rev

[g] GCN4-ILV1.2 = zeste-white

[h] GCN4-ILV1.2/Rev = zeste-white/Rev

[i] GR-uteroglobin.1 = GR-uteroglobin.2, GR/PR-MMTV.1, GR/PR-MMTV.2, GR/PR-MMTV.3, GR/PR-MMTV.4, GR/PR-MMTV.5

[j] GR-uteroglobin.1/Rev = GR-uteroglobin.2/Rev, GR/PR-MMTV.1/Rev, GR/PR-MMTV.2/Rev, GR/PR-MMTV.3/Rev, GR/PR-MMTV.4/Rev, GR/PR-MMTV.5/Rev

[k] GRE_CS7 = GRE_CS8/Rev

[l] hsp70.2 = Sp1-hsp70, Sp1-IE-3.3/Rev, Sp1-IE-4/5/Rev

[m] LVa-Mo-MuLV/Rev = LVa_RS/Rev

[n] rRNA-T2/T3 = T3-box

[o] SV40.13/Rev = SV40.16/Rev, SV40.6/Rev.

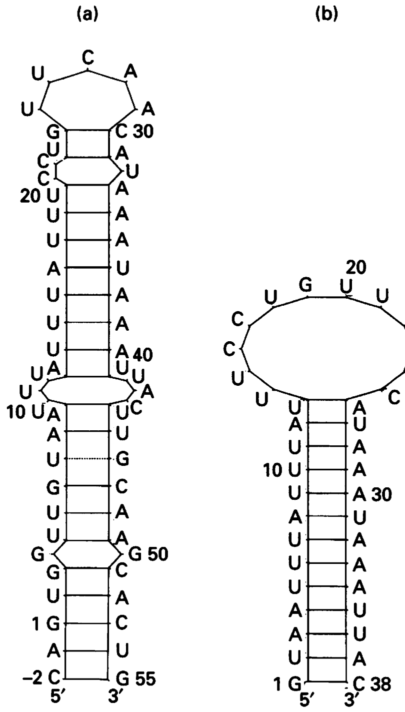

Figure 5 Stem loop structures of the 5′-end of human (a) and mouse (b) intron 2

The human sequence begins with the last two bases of exon 2 and includes 55 bases of intron 2. The mouse sequence begins with the first base of intron 2 and includes 38 bases of the intron.

pertain to cell-type-specific processing, since expression of the *COL2A1* gene in fibroblasts results in several unusually large mRNA species with longer 3′-non-translated sequences that have not been identified [38].

This work was supported in part by grants from the Medical Research Council of the Academy of Finland, and grants from the U.S. National Institutes of Health (AR39740), the Lucille P. Markey Charitable Trust, and the March of Dimes/Birth Defects Foundation.

REFERENCES

1 Bornstein, P., McKay, J., Morishima, J. K., Devarayalu, S. and Gelinas, R. E. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 8869–8873

2 Horton, W., Miyashita, T., Kohno, K., Hassell, J. R. and Yamada, Y. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 8864–8868

3 Rossi, P. and de Crombrugghe, B. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 5590–5594

4 Rossouw, C. M. S., Vergeer, W. P., du Plooy, S. J., Bernard, M. P., Ramirez, F. and de Wet, W. J. (1987) J. Biol. Chem. **262**, 15151–15157

5 Koltunow, A. M., Gregg, K. and Rogers, G. E. (1986) Nucleic Acids Res. **14**, 6375–6392

6 Ray, D., Meneceur, P., Tavitian, A. and Robert-Lezenes, J. (1987) Mol. Cell. Biol. **7**, 940–945

7 Moore, D. D., Marks, A. R., Buckley, D. I., Kapler, G., Payvar, P. and Goodman, H. M. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 699–702

8 Rieger, M. and Franke, W. W. (1988) J. Mol. Biol. **204**, 841–865

9 Upholt, W. B. (1989) in Collagen (Kang, A. H. and Nimmi, M. E., eds.), vol. 4, p. 31–49, CRC Press, Boca Raton, FL

10 Vuorio, E. and de Crombrugghe, B. (1990) Annu. Rev. Biochem. **59**, 837–872

11 Chu, M.-L. and Prockop, D. J. (1993) in Connective Tissue and its Heritable Disorders: Molecular Genetics and Medical Aspects (Royce, P. M. and Steinmann, B., eds.), pp. 149–165, Wiley–Liss, New York

12 Metsäranta, M., Toman, D., de Crombrugghe, B. and Vuorio, E. (1991) J. Biol. Chem. **266**, 16862–16869

13 Vikkula, M., Metsäranta, M., Syvänen, A.-C., Ala-Kokko, L., Vuorio, E. and Peltonen, L. (1992) Biochem. J. **285**, 287–294

14 Huang, M.-C., Seyer, J. M., Thompson, J. P., Spinella, D. G., Cheah, K. S. E. and Kang, A. H. (1991) Eur. J. Biochem. **195**, 593–600

15 Ryan, M. C., Sieraski, M. and Sandell, L. J. (1990) Genomics **8**, 41–48

16 Vikkula, M. and Peltonen, L. (1989) FEBS Lett. **250**, 171–174

17 Cheah, K. S. E., Stoker, N. G., Griffin, J. R. and Grosveld, F. G. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 2555–2559

18 Baldwin, C. T., Reginato, A. M., Smith, C., Jimenez, S. A. and Prockop, D. J. (1989) Biochem. J. **262**, 521–528

19 Elima, K., Vuorio, T. and Vuorio, E. (1987) Nucleic Acids Res. **15**, 9499–9504

20 Ala-Kokko, L. and Prockop, D. J. (1990) Matrix **10**, 279–284

21 Ala-Kokko, L. and Prockop, D. J. (1990) Genomics **8**, 454–460

22 Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. **12**, 387–395

23 Schuler, G. D., Altschul, S. F. and Lipman, D. J. (1991) Protein Structure Function Genet. **9**, 180–190

24 Upholt, W. B. and Sandell, L. J. (1986) Proc. Natl. Acad. Sci. U.S.A. **83**, 2325–2329

25 DePonti-Zilli, L., Seiler-Tuyns, A. and Paterson, B. M. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 1389–1393

26 Herget, T., Burba, M., Schmoll, M., Zimmermann, K. and Starzinski-Powitz, A. (1989) Mol. Cell. Biol. **9**, 2828–2836

27 Määttä, A., Bornstein, P. and Penttinen, R. P. K. (1991) FEBS Lett. **279**, 9–13

28 Pugh, C. W., Tan, C. C., Jones, R. W. and Ratcliffe, P. J. (1991) Proc. Natl. Acad. Sci. U.S.A. **88**, 10553–10557

29 Bodine, D. M. and Ley, T. J. (1987) EMBO J. **6**, 2997–3004

30 Lloyd, J. A., Krakowsky, J. M., Crable, S. C. and Lingrel, J. B. (1992) Mol. Cell. Biol. **12**, 1561–1567

31 Keffer, J., Probert, L., Cazlaris, H., Georgopoulos, S., Kaslaris, E., Kioussis, D. and Kollias, G. (1991) EMBO J. **10**, 4025–4031

32 Huff, A. C., Yuspa, S. H. and Rosenthal, D. (1993) J. Biol. Chem. **268**, 377–384

33 Ryan, M. C. and Sandell, L. J. (1990) J. Biol. Chem. **265**, 10334–10339

34 Eperon, L. P., Estibeiro, J. P. and Eperon, J. C. (1986) Nature (London) **324**, 280–282

35 Solnick, D. and Lee, S. I. (1987) Mol. Cell. Biol. **7**, 3194–3198

36 Watakabe, A., Inoue, K., Sakamoto, H. and Shimura, Y. (1989) Nucleic Acids Res. **17**, 8159–8169

37 d'Orval, B. C., d'Aubenton-Carafa, Y., Marie, J. and Brody, E. (1991) J. Mol. Biol. **221**, 837–856

38 Ala-Kokko, L., Hyland, J., Smith, C., Kivirikko, K. I., Jimenez, S. A. and Prockop, D. J. (1991) J. Biol. Chem. **266**, 14175–14178

Received 10 November 1994; accepted 30 January 1995