J Mol Evol (1985) 22:209–219

Journal of
Molecular Evolution

Evolution of Chick Type I Procollagen Genes

Kathy Benveniste-Schrode,[1] Jeffrey L. Doering,[2] Walter W. Hauck,[3] James Schrode,[4] Kari L. Kendra,[5] and Bonnie K. Drexler[5]

[1] Department of Oral Biology and [3]Biometry Section, Cancer Center, Northwestern University, Chicago, Illinois 60611, USA
[2] Department of Biology, Loyola University of Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626, USA
[4] Abbott Laboratories, North Chicago, Illinois 60064, USA
[5] Department of Biological Sciences, Northwestern University, Evanston, Illinois 60201, USA

Summary. Although the major types of vertebrate collagen have a number of structural properties in common, significant DNA sequence homologies have not been detected between different portions of the helical coding domains within the same gene or between different genes. However, under nonstringent hybridization conditions we found considerable cross-homology within and between α1(I) and α2(I) chick cDNAs in the coding regions for helical sequences. Detailed analyses at the DNA sequence level have led us to propose that the gene for chick pro α2(I) collagen arose from a 9-bp primordial sequence. A consensus sequence for the 9-bp repeat was derived: GGTCCTCCT, which codes for a Gly–Pro–Pro triplet. The primordial ancestor of this 9-bp unit, GGTCCTXCT, apparently underwent duplication and divergence. Each resulting 9-bp sequence was triplicated to form a 27-bp domain, and a condensation event produced a 54-bp domain. This genetic unit then underwent multiple rounds of amplification to form the ancestral gene for the full-length helical section of α2(I). A different 9-bp consensus sequence (GGTCCCCCC) seems to have been the basis of the chick pro α1(I) gene.

Key words: Collagen — Ancestral genetic domain — Simple sequence amplification — Consensus sequence — Introns

Introduction

The collagens are a family of proteins that constitute the major component of the extracellular matrix in most animal tissue. In higher vertebrates there are three major interstitial collagens, with four genes required to code for their various constituent chains, which are synthesized as procollagen precursors (Bornstein and Sage 1980). These collagens share several structural properties, including a glycine residue at every third position in the helical domains of the polypeptide chains, a high proline and hydroxyproline content, and three polypeptide chains that interact to form a helical molecule (Bornstein and Sage 1980; Eyre 1980). The procollagen precursors have extensions at both the amino and carboxyl terminals. The former contains one short collagenous helical section, while the latter contains none (Bornstein and Sage 1980).

Overall, the sequence of the helical region has remained highly conserved both within the collagen gene family and from species to species (Hofmann et al. 1980). Several investigators have examined the amino acid sequences within helical regions looking for repeat patterns (Piez and Torchia 1975; McLachlan 1976; Trus and Piez 1976; Hofmann et al. 1978, 1980). The distribution of charged and hydrophobic residues shows a periodicity of 234 residues (Piez and Torchia 1975; McLachlan 1976), the same as the length of the D-stagger of collagen fibrils. On the basis of the D repeat structure, McLachlan (1976) suggested that collagen sequences evolved from a primordial gene of length D by gene duplication. When Hofmann et al. (1980) extended the comparison of internal sequences to include interactive and noninteractive residues, repeat units 78, 39, 21, and 18 amino acids in length were observed. They proposed that the 78-residue repeats, which are defined mainly by noninteractive residues, may better reflect the original genetic unit, since such residues have not been exposed to selective pressure directed toward improved aggregation of the molecules into fibrils. Therefore, they con-

Offprint requests to: J.L. Doering

cluded that the 78-residue unit was the primordial genetic unit. However, they did not preclude the possibility that this sequence arose by duplication of a 39-residue unit that in turn consisted of an 18- and a 21-residue unit.

The availability of DNA sequence data affords us a better opportunity for examining the question of collagen gene evolution. Fuller and Boedtker (1981) and Monson and McCarthy (1981) have published the partial sequences of α1(I) and α2(I) cDNAs from chick and mouse. Their results support the observation, made by direct amino acid analysis (Hofmann et al. 1980), that the helical domain is quite conserved across species lines. In addition, Tate et al. (1983) and Bernard et al. (1983) have found a short region in the carboxyl-terminal propeptide that is highly conserved in amino acid and DNA sequence among the various collagen genes and among species. Codon usage analyses have shown a preference for U and C in the third positions of glycine, proline, and alanine codons (Fuller and Boedtker 1981; Bernard et al. 1983). However, although the tripeptide repeat structure has been recognized as elemental to the helical nature of collagen, there has been no analysis of DNA sequences that would support the hypothesis that a nine-nucleotide sequence served as the primordial gene sequence.

Instead, two hypotheses concerning the evolution of the collagen gene have emerged, based on the presumed existence of an ancestral gene 54 bp long that codes for a peptide 18 amino acids long. The development of these hypotheses rests on the observations that (1) the vertebrate collagen genes studied so far are many times longer than their corresponding mRNAs because of the presence of approximately 50 introns of varying lengths, and (2) while all the exons are multiples of 9 bp, many are multiples of 54 bp (Yamada et al. 1980; Monson and McCarthy 1981; Wozney et al. 1981).

The first hypothesis suggests that the full-length collagen gene arose by multiple duplications of a single genetic unit containing 54 bp and the *adjacent* sequences required for splicing (Yamada et al. 1980). The second hypothesis suggests that the primordial collagen gene arose by a series of homologous recombinational events *within* coding sequences and that the intervening sequences were inserted at a later time (Monson and McCarthy 1981; Wozney et al. 1981). Support for this hypothesis is taken from the large number of exons that are not 54 bp or multiples thereof and the homology seen between exon sequences in staggered alignment. In contrast to previous studies, our analysis specifically examines the available DNA sequences for chick α1(I) and α2(I) collagen gene coding regions in search of a primordial sequence without any bias based on exon size. In the case of the α2(I) gene, the results strongly suggest the existence of a nine-nucleotide primordial gene (GGTCCTXCT) that underwent duplication and divergence. Each resulting 9-bp sequence was then triplicated to form a 27-bp domain. A condensation event produced a 54-bp domain, and this genetic unit then underwent multiple rounds of amplification to form the ancestral gene that codes for the full-length helical section of α2(I). A different 9-bp consensus sequence (GGTCCCCCC) seems to have been the basis of the chick pro α1(I) gene.

Materials and Methods

Computer-Assisted Searches. Computer programs were written in TRS-80 BASIC for searching the DNA sequences for direct repeats of preset lengths. The available DNA sequences for chick α1(I) and α2(I) gene coding regions (Yamada et al. 1980; Fuller and Boedtker 1981; Wozney et al. 1981; Tate et al. 1983) were entered into the programs. Where both cDNA and genomic sequences are known, no differences have been observed. Where available, the genomic sequences were used; otherwise cDNA sequences were entered into the program. The data for the α2(I) chain represent 18 of the 40 exons that contain sequences for the major triple-helical domain. These exons are spread throughout the gene, together representing 44% of the coding sequence for the helix.

Determination of the DNA Consensus Sequence. The sequence for each chain was listed in groups of nucleotides of the desired length (9, 18, 27, and 54). For 9-, 18-, and 27-bp lengths all available sequences were analyzed as if they were contiguous, without regard to location within the helical coding domain or exon organization. Position 1 was assigned to the first base of the first glycine codon. For the 54-bp lengths we used only sequences to which positions in an exon had been assigned. For exons shorter than 54 bp, the sequence was started in position 1 and spaces were placed at the end of the exon; for exons longer than 54 bp, the sequence was divided into sections a and b and if the exon was shorter than 108 bp, spaces were placed at the end of section b. The numbers of times the bases appeared in each position were tabulated, and the most frequently appearing base in each position defined the consensus sequence.

Hybridizations and Melting Curves. Restriction digests of the inserts from pCg54 or pCg45 (Fuller and Boedtker 1981) were run on 2.5% agarose gels as previously described (Peterson et al. 1980) and transferred to nitrocellulose filters by the method of Southern (1975). Clones containing the plasmids were generously provided by H. Boedtker. Probes were ^{32}P-labeled uniformly by nick translation or at 5′ termini by T4 polynucleotide kinase (Peterson et al. 1980). Filter hybridizations and subsequent washes were done according to our previously described methods (Doering et al. 1982), except that hybridizations were done at 50°C without formamide, and filters were washed for 2 h at 50°C in eight changes of hybridization solution, followed by four 15-min washes at room temperature in melt buffer (50 mM NaCl; 10 mM Tris-HCl, pH 7.5; 1 mM ethylenediaminetetraacetate). Areas of a filter containing fragments to which a probe had hybridized were cut out and incubated for 10 min in melt buffer at temperatures increased by 5°C intervals. The radioactivity remaining bound to the filter at each temperature was measured, permitting the construction of melting curves. The T_m for a filter-bound probe homoduplex was compared with those for hybrids

of the probe with heterologous fragments. Such filter hybridizations give a qualitative indication of which fragments share sequence similarity, but cannot be used to quantitate accurately the extent of that similarity.

Statistical Methods. Our statistical analyses are based on various chi-square techniques. In the analysis of the number of mismatches, each consecutive 9-bp portion of a DNA sequence was compared with its appropriate consensus sequence and the number of mismatches determined. The respective numbers of 9-bp sequences with zero to nine mismatches were tallied. We then used a chi-square goodness-of-fit test to compare this distribution with that expected assuming that the replicated consensus sequence had undergone random mutation by a binomial model. A statistically significantly large chi-square value is indicative of nonbinomial mutation.

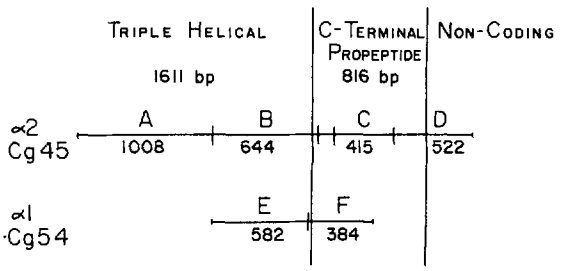
Our other analyses are based on log-linear analysis (Fienberg 1977). This is a method for analyzing tables of counts where the dimension of the table is greater than two. Log-linear analysis can be viewed as a generalization of the usual chi-square techniques that are used for two-dimensional tables.

In our analyses of patterns of number of mismatches we looked at how the distribution of number of mismatches varied with number of mismatches in the previous 9-bp unit. This was a two-way analysis for α1(I), where only the DNA sequence from the region corresponding to amino acids 814–1014 was used; for α2(I) we also included the four sequence areas arbitrarily assigned as the four regions of known DNA sequence (amino acids 37–72, 175–360, 379–429, and 814–1014) as a factor.

For our analysis of a potential 54-bp consensus sequence in α2(I) we looked at how the distribution of bases varied with location within the 9-bp unit and across replicate 9-bp units. Only the seven variable locations in the 9-bp unit (positions 3–9) were used in these analyses (see Table 1). We then divided the 9-bp sequences into two groups of three sequences each comprising the first 27 and the last 27 positions of the 54-bp sequence. These analyses were done for the seven variable locations (3–9) and then separately for the four C positions (4, 5, 7, and 8) and two T positions (6 and 9) of the 9-bp consensus sequence (see Table 1). The results from these three analyses were the same; we report the results using all seven variable locations. Similar analyses were done for potential 9-, 18-, and 27-bp consensus sequences.

Results

We determined the extent of cross-hybridization between chick α1(I) and α2(I) cDNAs (Fig. 1) in the course of developing conditions under which they could be used as specific probes in other species for only their isologous genes. Low-stringency hybridization (see Materials and Methods) reproducibly detected significant hybridization between these two sequences. Lambda phage DNA did not hybridize to either cDNA under these conditions. Thus, nonspecific adsorption of the probe could not account for these results. The regions of cross-homology between the two sequences were localized using various restriction fragments (Fig. 1). At a qualitative level there is no significant sequence homology between α1(I) and α2(I) in the carboxy-terminal propeptide coding regions or between the triple-helical region and carboxy-terminal propeptide region



Fragments Hybridized	T_M Reduction (°C)
A x E	20
B x E	20
C x E	no hybrid
D x E	no hybrid
B x A	15
C x A	no hybrid
D x A	no hybrid
F x A	no hybrid

Fig. 1. Localization of the regions of cross-similarity in the α1(I) and α2(I) chick collagen cDNAs. Hybridizations involving the indicated restriction fragments were performed as described in Materials and Methods. Fragment sizes are given in base pairs (bp). The T_m reductions were determined as described in Materials and Methods

within a cDNA. The α1(I) cDNA clone used here does not extend into the short region of sequence similarity between α1(I) and α2(I) carboxy propeptide coding regions (Bernard et al. 1983; Tate et al. 1983). However, clearly there is significant similarity between α1(I) and α2(I) sequences in the triple-helical coding regions that has not been previously noted. Moreover, there is a good degree of similarity between the two fragments within the triple-helical coding region of the α2(I) cDNA (B × A). This finding was not predicted from prior analyses of the α2(I) cDNA sequences (Fuller and Boedtker 1981) and suggests the presence of a previously undescribed tandemly repeated sequence.

Computer-assisted searches of the α1(I) or α2(I) collagen coding sequences showed no direct repeats longer than 17 bp and very few longer than 12 bp. Taken together with the tripeptide repeating amino acid pattern seen in the helical region of collagen, these data suggested a nine-nucleotide sequence as the most likely basis for any tandem repeat.

When the published nucleic acid sequence for the triple-helical region of chick α2(I) collagen was written as 9-bp units and the frequencies of the bases in each position were tallied (see Materials and Methods), a consensus sequence was identified (Table 1). The frequencies of 1.0 in the first two positions are to be expected because of the requirement for glycine as every third amino acid of the helix and because the codons for glycine are of the form GGX. These then are the invariable positions,

Table 1. Primordial consensus sequences for chick α1(I) and α2(I) collagen genes

Base	Position 1	2	3	4	5	6	7	8	9
				Chick α1(I) DNA sequence					
G	67	67	1	23	3	5	19	9	8
C	—	—	24	39	43	35	32	40	38
A	—	—	4	3	15	9	13	13	6
T	—	—	38	2	6	18	3	5	15
Consensus:	G	G	T	C	C	C	C	C	C
	1.00	1.00	0.57	0.58	0.64	0.52	0.48	0.60	0.57
				Chick α2(I) DNA sequence					
G	158	158	4	43	9	17	40	25	16
C	—	—	33	80	78	35	69	77	6
A	—	—	22	24	45	31	44	37	22
T	—	—	99	11	26	75	5	19	114
Consensus:	G	G	T	C	C	T	C	C	T
	1.00	1.00	0.63	0.51	0.49	0.47	0.44	0.49	0.72

The available DNA sequences for the helical domains of these collagen chains (Yamada et al. 1980; Fuller and Boedtker 1981; Wozney et al. 1981) were entered into a TRS-80 BASIC program. The sequence for each chain was listed in consecutive groups of nine nucleotides and the consensus sequences were derived as described in Materials and Methods. At the bottom of each section, the base that appears most often in each position of the 9-bp sequence is indicated along with its frequency

Table 2. Frequency distributions of numbers of mismatches for 9-bp sequences compared with the appropriate consensus sequence

No. of mismatches	Gene α1(I)	α2(I)
0	1 (1.5%)	4 (2.5%)
1	13 (19.4%)	24 (15.2%)
2	16 (23.9%)	33 (20.9%)
3	12 (17.9%)	29 (18.4%)
4	10 (14.9%)	28 (17.7%)
5	9 (13.4%)	22 (13.9%)
6	4 (6.0%)	15 (9.5%)
7	2 (3.0%)	3 (1.0%)
8	0	0
9	0	0

Each consecutive 9-bp portion of DNA sequence was compared with its appropriate consensus sequence and the number of mismatches determined. Nine-base-pair sequences with zero to nine mismatches are tallied here. The numbers in parentheses represent the distributions expressed as percentages

whereas the remaining seven positions of the 9-bp unit are potentially variable. The frequent appearance of specific nucleotides at each potentially variable position has led us to conclude that the sequence GGTCCTCCT is a consensus sequence for the chick α2(I) gene. This sequence codes for a Gly–Pro–Pro tripeptide.

Each 9-bp unit of the α2(I) DNA sequence then was compared with the consensus sequence and the number of mismatches was determined. Nine-base-pair units with zero to seven mismatches were tallied and the distribution was calculated on a percentage basis (Table 2). A chi-square goodness-of-fit test compared this distribution with that expected assuming that the replicated consensus sequence had undergone random mutation by a binomial model. A large chi-square value is indicative of nonrandom substitution. The mismatch distribution is significantly skewed toward low numbers of mismatches and indicates that the α2(I) sequence has not undergone binomial mutation (chi-square = 42.12 on 6 degrees of freedom, $P < 0.001$). This argues that the consensus is good and that there has been evolutionary pressure to maintain this sequence.

The α2(I) sequences analyzed come from a number of regions spread throughout the gene, together totalling 44% of the coding sequence for the helix. A histogram of the number of mismatches between each nine-nucleotide sequence and the consensus sequence for α2(I) (Fig. 2) reveals no large regions of high conservation or divergence from the consensus sequence. Rather, 9-bp units with two or less mismatches, that is, those with more than 70% matches in the variable bases, occur uniformly throughout the sequence (Fig. 2); no statistically significant association with the number of mismatches in the preceding 9-bp unit was detected ($P = 0.873$; see Materials and Methods). This all tends to indicate that the 9-bp consensus sequence we have determined is the primary structural unit for the *entire* α2(I) gene's helical coding domain.

The 9-bp repeat exhibits an unusual polarity in the conservation of nucleotides at each position. Bases are more conserved at the first and last variable positions than in the middle (Table 1). The strong conservation in these positions is especially

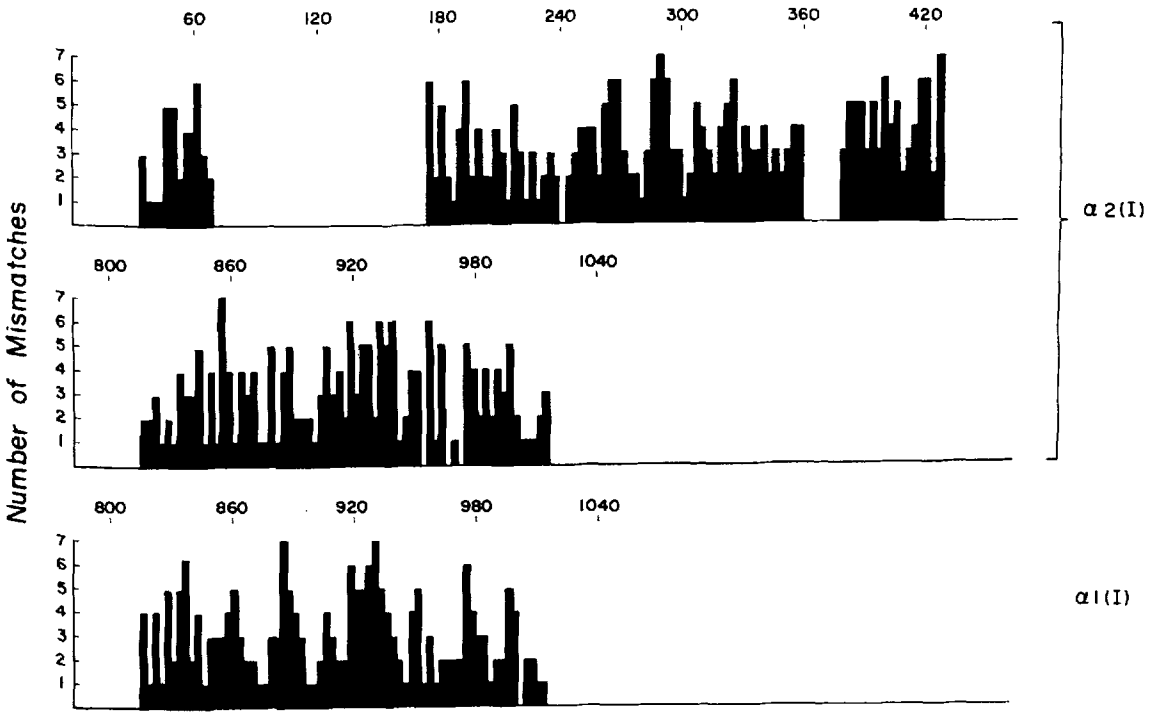


Fig. 2. Histogram of the deviation from the consensus sequence of every 9-bp unit within chick α1(I) and α2(I). The DNA sequences for the helical regions of each of these collagen chains were compared with their respective consensus sequence and the number of mismatches was plotted as a function of position in the helix. The amino acid positions of the 9-bp units are indicated on the abscissa. The gaps in the α2(I) histogram are regions where DNA sequence is not available

striking considering that variations in these positions would evoke a silent mutation in the wobble position of the Gly and Pro codons. In contrast, the first position in a codon, 7, has the lowest degree of conservation, with both G and A occurring nearly as frequently as C. Subsequent analysis (see below) showed that this may well be the result of the process that led to the evolution of a full collagen gene.

Attempts were next made to determine whether a repeating sequence larger than 9 bp could have served as the primordial element in the collagen gene's formation. One of the most logical sizes to consider is 54 bp, since many exons are this size or a multiple thereof. On the basis of these observations, Yamada et al. (1980) suggested that the 54-bp exons represent a primordial sequence. In the current case, we analyzed only sequences for which position in an exon had been assigned. This limited the analysis to α2(I), but excluded only a small percentage (72 of 1422 bp) of the data used in our previous analyses. Table 3 shows the consensus sequences and the frequencies of the bases if repeating units of 9, 18, 27, or 54 bases are assumed. The average frequency for the variable bases remains quite constant independent of the size of the repeating unit. The 18- and 27-bp consensus sequences give no strong evidence of being anything other than multiple tandem copies of the fundamental 9-bp unit. However, the 54-bp consensus sequence is *not* composed of six identical copies of the 9-bp unit. The percentage of matches of each 54-bp unit with the consensus sequence, in the variable positions only, is shown in Table 4. More than 75% of the repeats have more than 50% matches with the consensus sequence, the greatest similarity being 71%, for exon 4. The one repeat with less than 40% matches is the first 54 bp of exon 7, which contains 33% matches. This is of special interest, since this is a sequence that codes for a cross-linking region (lysine in position 937 of the helix) (Bornstein and Sage 1980), and thus may have different constraints than the rest of the helical coding region. These results argue that the 54-bp consensus sequence is a good one that is significantly different from just a tandem array of homogeneous 9-bp units. The simplest assumption is that the 9-bp unit was the primordial sequence from which the gene arose, and that a 54-bp unit was also a discrete step in the gene's evolution.

We thus wanted to examine the various consensus sequences in Table 3 to see whether there was any statistical evidence for the 9-bp unit having been amplified to form some larger unit that in turn served as an intermediate in collagen gene construction. To do this we used log-linear analysis, assessing the goodness of fit of various models by likelihood ratio

Table 3. Consensus sequences derived with various repeating lengths

Consensus sequence																			
9-Mer																			
G	G	T	C	C	T	C	C	T											
		0.63	0.51	0.49	0.47	0.44	0.49	0.72											
18-Mer																			
G	G	T	C	C	T	C	C	T		G	G	T	C	C	T	C	C	T	
		0.71	0.50	0.49	0.42	0.49	0.58	0.78				0.51	0.53	0.50	0.51	0.39	0.40	0.69	
				0.57										0.50					
27-Mer																			
G	G	T	C	C	T	C	C	T		G	G	T	C	C	T	C/G	C	T	
		0.56	0.50	0.48	0.42	0.50	0.44	0.73				0.54	0.54	0.54	0.54	0.38	0.52	0.77	
				0.52										0.55					
54-Mer																			
G	G	T	C	A	T	C	C	T		G	G	C	C	C	T	C	C	T	
		0.62	0.50	0.46	0.38	0.69	0.62	0.85				0.42	0.54	0.54	0.62	0.50	0.54	0.92	
				0.59										0.58					
G	G	T	C	C	T	A	A	T		G	G	T	C	C	T	G	C	T	
		0.50	0.50	0.58	0.46	0.42	0.31	0.62				0.69	0.54	0.54	0.46	0.42	0.50	0.62	
				0.48										0.54					

The available coding sequence for the chick $\alpha 2$(I) gene (Yamada et al. 1980; Fuller and Boedtker 1981; Wozney et al. 1981; Tate et al. 1983) was listed in groups of 9, 18, 27, or 54 bp. The number of times the bases appeared in each position was tabulated and consensus sequences were determined as described in Materials and Methods. Where the base repeat is larger than nine bp the average frequency for each 9-bp subunit is shown in the line below the individual base frequencies. The overall average frequency for the full-length repeat is shown on the right

tests. This method takes the overall ("common") distribution of bases at each of the seven variable locations and asks whether each subregion is statistically consistent with this distribution, and if not, where the differences occur. This analysis does not assume a random base pair distribution and thus is particularly helpful in analyzing nonrandom coding sequences like collagen.

The consensus sequence for the hypothetical 18-bp unit (Table 3) gives no indication of being anything other than two tandem 9-bp units. This observation was substantiated by the log-linear analysis, which found no significant difference in the base pair distributions of the two 9-bp units ($P = 0.471$); that is, the base distributions at each of the seven variable locations of the two 9-bp units were consistent.

If we instead assume a 27-bp consensus sequence, the three 9-bp units do not differ significantly ($P = 0.095$). This result is closer to indicating a statistically significant difference than that for the 18-bp consensus sequence, reflecting the nonhomogeneity of the three 9-bp units within the consensus sequence; in location 7, C is a fairly strong consensus for the first and third 9-bp unit, while in the middle 9-bp unit G is only a weak consensus (20 Gs and 19 Cs). We also divided the 27-bp sequence into a 9-bp unit and an 18-bp unit consisting of two homogeneous 9-bp subunits. When we take the 9-bp unit as the start of the 27-bp sequence, we find that the 18-bp unit and 9-bp unit are homogeneous ($P = 0.235$). However, when we take the 9-bp unit as the end of the 27-bp sequence, we find marginal heterogeneity ($P = 0.040$). Together, the 27-bp consensus sequence results hint at some underyling heterogeneity, but are not conclusive.

We analyzed the 54-bp consensus sequence as two 27-bp units, as three 18-bp units, and as six 9-bp units. Beginning with the 27-bp unit analyses, we first noted heterogeneity between the two halves of the 54-bp sequence (Table 3); there is disagreement in the consensus base at 6 of the 27 locations, three of which (7, 16, and 25) correspond to the same relative location within a 9-bp subunit. The difference between the two 27-bp subsequences was most notable when positions 7, 8, and 9 were compared with 34, 35, and 36. For example, at position 7, there are 45 Cs and 14 As among the 78 bases in the first 27-bp sequence; in the analogous position of the second 27-bp sequence, position 34, there are 21 Cs and 28 As among the 72 bases. This is the basis for the low degree of sequence conservation at

Table 3. Extended

									Average frequency
									0.53
									0.53
G	G	T	C	C	T	C	C	T	
		0.76	0.50	0.46	0.43	0.46	0.52	0.65	
				0.54					0.54
G	G	T	C	C	T	C	C	T	
		0.81	0.46	0.54	0.42	0.54	0.62	0.77	
				0.59					
G	G	T	C	C/T	T	A	C	T	
		0.70	0.55	0.35	0.45	0.45	0.40	0.50	
				0.49					0.55

position 7 seen when the sequence is analyzed as a single 9-bp repeat (Table 1, and see above). For the log-linear analyses, we considered each 27-bp unit as consisting of three 9-bp subunits so that we could look for heterogeneity both between and within the 27-bp units. We found that the base pair distribution differed significantly between the two 27-bp units ($P = 0.004$), but that there was no significant heterogeneity within each of the 27-bp units ($P = 0.373$). The 9-bp consensus is GGTCCTCCT for the first 27-bp unit and GGTCCTACT for the second 27-bp unit. These observations strongly suggest that a 27-bp domain was an intermediate between the 9-bp and a 54-bp stage in the gene's evolution.

In the three 18-bp unit analysis, the three 18-bp units just miss being statistically significantly heterogeneous at the 5% level ($P = 0.051$). Again there is no evidence for within-unit heterogeneity ($P = 0.186$), but the lower P-value compared with that in the 27-bp unit analysis (0.186 compared with 0.373) indicates that two homogeneous 27-bp units give a better fit to the data than three homogeneous 18-bp units. The 9-bp consensus for the third of the three 18-bp units is the same as for the second 27-bp unit. The better fit of the two 27-bp unit model suggests that the fourth 9-bp subunit belongs with the last two subunits, not with the first three as in the three 18-bp unit analysis.

When the 54-bp consensus sequence was analyzed as six 9-bp units, the base pair distribution was found to differ significantly between units ($P = 0.032$). Thus, there is no evidence for the 54-bp sequence being composed of six homogeneous 9-bp units.

Altogether, our log-linear analyses suggest that the gene did not evolve directly by multiple rounds of amplification of the 9-bp unit, but that subsequences of 27 and 54 bp served as intermediates in the gene's construction. The data also indicate that the first 27 bases of the 54-bp domain were derived from one 9-bp unit, GGTCCTCCT, and the second 27 bases were derived from another 9-bp unit, GGTCCTACT. These sequences were determined from the consensuses of the three 9-bp units in the first and second 27-bp domains, respectively (Table 3). A model for the evolution of collagen gene structure that incorporates all these elements is presented in the Discussion.

Some of the analyses described above were also done on the chick $\alpha 1$(I) DNA sequence (Table 1). As with the $\alpha 2$(I) sequences, a consensus sequence became apparent: GGTCCCCCC. While this sequence still codes for a Gly–Pro–Pro triplet, it is distinct from that derived for the $\alpha 2$(I) chain. There is a lower consensus frequency at positions 3 and 9 and a higher consensus frequency at positions 4, 5, 6, 7, and 8 than in $\alpha 2$(I). Thus, there is a shift to lower mismatch frequencies in $\alpha 1$(I) (Table 2). A goodness-of-fit test indicates that mutation away from the consensus has produced fewer changes than would be predicted by random binomial mutations (chi-square = 15.31 on 6 degrees of freedom, $P = 0.018$). Thus, as for $\alpha 2$(I), $\alpha 1$(I) has a good consensus sequence that seems under some pressure to be maintained.

When the pattern of mismatches in $\alpha 1$(I) was analyzed by the log-linear analysis it was found to be nonrandom in the sense that the number of mismatches in a 9-bp unit was related to the number of mismatches in the previous unit ($P = 0.053$); 9-bp units in good agreement with the consensus are in clusters and those in poor agreement are in other clusters (Fig. 2). Further analysis of $\alpha 1$(I) was limited by the amount of sequence data available.

Discussion

The high degree of similarity detected between the two fragments within the triple-helical coding region of the $\alpha 2$(I) cDNA was not predicted from prior analyses of the $\alpha 2$(I) cDNA sequences (Yamada et

Table 4. Homology between chick α2(I) DNA sequences and 54-bp consensus sequence

Exon	Consensus: GGTCATCCTGGCCCTCCTGGTCCTCCTGGTCCTAATGGTCCTGCTGGTCCTACT	% Matches
41	CC AA T G C CCA AAAG GAAGA	55
40	C AAA AAGA GAG GG GT CAA	57
34a	G AATC A G AA GAA C C A GA	60
34b	AG GAT A T T CAG GT C C C----------	54
33	CA C TG AA T GG A G A ----------	63
32a	T TGT G G TGCC G CCG AT C	60
32b	C G AAG G GG A T T ----------	69
31	G A A G G CAAG AGAA G AACAAG GAGC	43
30a	GC G C C A G GAG AA CAAG GA	57
30b	CAGCAA TGAA CT G CCC C TA GA	50
29	CG G AT G TC AG G CAGA GTC TG	50
28a	A C G TAAC G G AG A GT G AAG A	55
28b	G G G GAA T TG AAGA----------	63
26a	TTC AG AGA AGGGT G A TC A C AA GA	43
26b	G A AACAT ATTC A A A CA ----------	54
25	G G AAA GAAAAA CAA GTC T C ACGG	45
10	CA AAA GA A C G G G C CTTC AG	55
9a	G G G TG AAC A C G G TTG G C	60
9b	C A AAGT T AAAGCC AAACCG GA C	48
8	C GT T GT G G TT C AAGA T G	60
7a	C CA AA T A G GAGAAA GAACC GA AAG A A GA	33
7b	TG TGAAG A ACAA ATTGC G T C T G	48
6	C A A TGA AA AAC C C A GG	62
5a	C T T AAGG GCAA TCC	67
5b	A CCAT G GTA G AT C AGCCAA C G	50
4	C C C C CCC CAA ----------	71

The DNA sequences of chick α2(I) that have been assigned an exon position (Yamada et al. 1980; Fuller and Boedtker 1981; Wozney et al. 1981; Tate et al. 1983) were listed by exon in lengths of 54 bp. The consensus sequence was derived as described in Materials and Methods. Sites that differ from the consensus sequence are shown; dashes indicate positions missing in exons shorter than 54 or 108 bp. On the left margin each sequence is identified by its exon number (Tate et al. 1983) and on the right margin the percentage of matches with the consensus, in the variable positions only, is listed

al. 1980; Fuller and Boedtker 1981) and suggests the presence of a previously undescribed, tandemly repeated sequence. The repeating structure of collagen, in which glycine occurs as every third amino acid, suggested that a nine-nucleotide sequence might well be the basis for any tandem repeat. Consensus sequences of nine nucleotides were identified for the α1(I) and the α2(I) chains (GGTCCCCCC and GGTCCTCCT, respectively). Both these sequences code for a Gly–Pro–Pro peptide. A tripeptide with just such a sequence is a likely candidate for the basic building unit of collagen, since after glycine, proline (proline plus hydroxyproline) is the second most abundant amino acid in collagen. Nine-base-pair units with more than 70% matches with the appropriate consensus sequence appear throughout the α1(I) and α2(I) sequences, often in tandem. Therefore, it is most likely that the consensus sequence represents a primordial sequence that was amplified many times to produce the genetic domain coding for the helical region of the ancestral collagen gene.

Previous studies (Yamada et al. 1980; Wozney et al. 1981) emphasized that many exons in the chick α2(I) gene are 54 bp or direct multiples thereof long, thus prompting the suggestion that a 54-bp sequence served as the primordial gene. Our clear demonstrations of a 9-bp consensus sequence throughout the gene's coding regions and the existence of discrete 27-bp domains indicate that there were at least two stages in the evolution of the gene prior to the appearance of the 54-bp domain.

Our analyses have suggested a detailed model for the evolution of the chick α2(I) gene, shown in Fig. 3. The primordial 9-bp sequence, GGTCCTXCT, duplicated and the two copies diverged from each other to form the closely related sequences GGTCCTCCT and GGTCCTACT. Separated from each other, each of these sequences was then triplicated to form the two 27-bp domains. Removal of the sequences between these two domains (condensation) resulted in the appearance of the ancestral 54-bp domain. This sequence then mutated to form the 54-bp concensus sequence, which in turn underwent multiple rounds of amplification to form the full-length helical coding region.

There are several mechanisms by which the 54-bp amplification could have occurred. One previ-

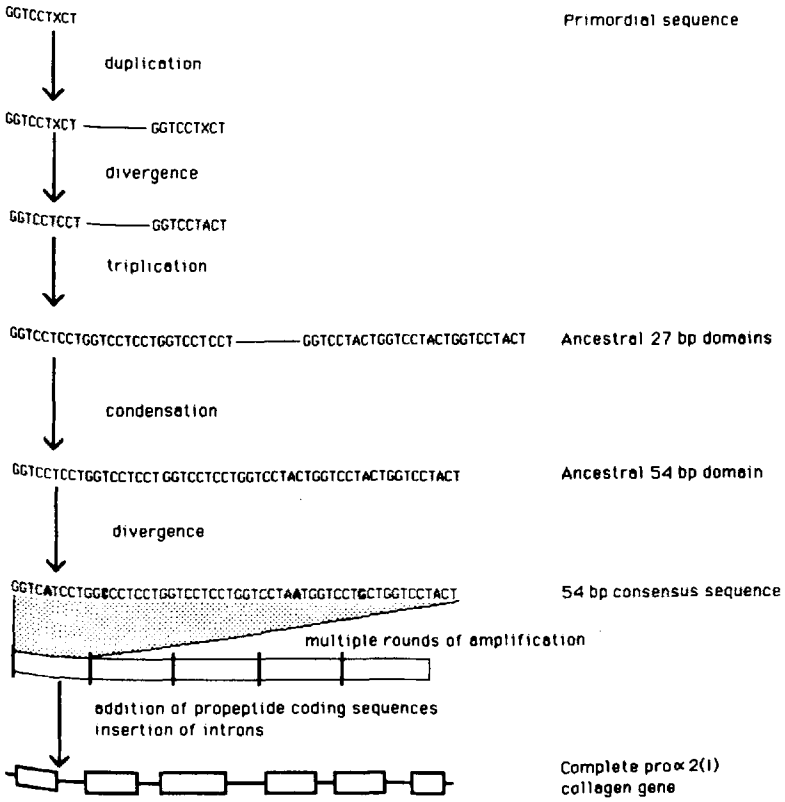


Fig. 3. A scheme for the evolutionary development of vertebrate collagen genes. The data presented in this article suggest that a primordial nine-nucleotide sequence underwent tandem duplication to form two independent genetic domains. These diverged and then triplicated to form two 27-bp domains. A condensation event joined these two domains to form a 54-bp domain. The 54-bp sequence underwent a number of tandem duplications to form the full-length triple-helical domain of collagen. Propeptide coding regions were then added, followed by the insertion of intervening sequences (introns) to form a complete collagen gene transcription unit. The various collagen genes could then have evolved from this original transcription unit after it had duplicated several times. See the text for further justification of this scheme

ously suggested possibility is that a unit consisting of the 54 bp plus adjacent sequences required for RNA splicing gave rise to the full-length gene by multiple duplications involving recombination between introns (Yamada et al. 1980). Alternatively, multiple rounds of homologous recombination within the 54-bp unit could have resulted in a full-length helical coding region, with intervening sequences being inserted at a later time (Monson and McCarthy 1981; Wozney et al. 1981).

We favor a model in which the introns were inserted after the 54-bp unit had been amplified to form the complete helical coding region (Fig. 3). The highly repetitious nature of the 54-bp unit, with its 9-bp and 27-bp subunits, makes it very likely that the gene expanded in size by a mechanism such as out-of-register homologous recombination (Smith 1976), which does not require introns as a part of the amplification event. The presence, near the 5′ and 3′ ends of the gene, of "junction" exons, which contain both helical and propeptide coding regions (Tate et al. 1983), could also argue that introns were added late, after helical and propeptide regions had been assembled to form a full gene. Our preliminary analysis of the collagen amino acid sequence has identified a 30-amino-acid repeating unit that, at the gene level, is interrupted by introns, which further suggests that introns were inserted relatively late in the gene's evolution. Finally, nonfibrous collagen genes in insects (Monson et al. 1982) and nematodes (Kramer et al. 1982) have very few introns, again implying that the numerous introns in vertebrate fibrous collagen genes (Tate et al. 1983) arose in a more recent evolutionary event.

The limited number of exon sizes in the α2(I) gene (Tate et al. 1983) would imply that intron insertion was not a random event. While previous studies stressed the numerous 54-bp exons, in fact the majority of the helical region coding information is in exons of other sizes (Tate et al. 1983). Thus, introns would not have to have been inserted at intervals as precisely as was earlier supposed. Recent studies indicate that only a small number of discrete exon sizes are found in all higher eucaryotic genes (Naora and Deacon 1982), so the limited variation in α2(I) exon size is less surprising than it seemed. There may also have been some selective pressure to place introns at particuilar sites in collagen genes to meet the requirements for processing such a long (>38 kb), internally repetitious gene transcript.

Intron insertion would likely have reduced the extent of recombination within the repetitious gene sequence. Thus, there could have been a selective advantage to adding the large number of introns, as it would have stabilized the size and structure of the collagen molecule. Indeed, genes with internally repetitious structures but few introns show a good deal of instability in their structures (Manning and Gage 1980).

That introns are located at analogous positions in all vertebrate fibrous collagen genes so far examined (Tate et al. 1983; Chu et al. 1984; Yamada et al. 1984) suggests that introns were inserted into an ancestral collagen gene prior to duplication of the entire gene. Presumably, multiple rounds of such duplication followed by sequence divergence led to the evoluton of the various types of interstitial collagen genes. We have been able to assign sequences to each of the apparent subunit stages in α2(I) gene construction, and so for clarity our model (Fig. 3) is presented with α2(I) as the ancestral complete gene. The lack of sufficient DNA sequence data for other collagen genes at this time prevents a determination of the actual ancestral gene sequence. Nevertheless, whatever its sequence, our work indicates that the ancestral gene was formed by steps analogous to those we have proposed in Fig. 3.

The internally highly repetitious nature of collagen genes would permit a mutation to be spread rapidly to every repeating unit throughout the gene's length. The presence of introns would not serve as an impediment to gene conversion events (Slightom et al. 1980; Baltimore 1981), which are believed to occur much more frequently than homologous recombination (Baltimore 1981). Thus, once multiple copies of the complete collagen gene existed, each copy could undergo its own intragenic conversion events (rectification) completely independently of other members in the collagen gene family. This could account for the difference in the 9-bp unit consensus sequences we observed for the α1(I) and α2(I) genes. That the 54-bp consensus sequence is marginally better than consensus sequences of other lengths in the α2(I) gene (Table 3) could also be due to relatively recent conversion events between exons in the gene. Intragenic rectification might also explain the dramatically different codon usages in the different collagen genes (Fuller and Boedtker 1981; Bernard et al. 1983; Tate et al. 1983). Independent rectification within each gene could likewise account for the surprisingly long divergence time found between the α1(I) and α2(I) genes when the estimate assumes a uniform rate of mutation (Bernard et al. 1983). Intragenic recombination may be the cause of the deletions that occasionally occur in collagen genes (Chu et al. 1983).

Within the last few years, several new collagenous molecules have been reported that have a major helical domain that is either shorter or longer than that present in the interstitial collagens (Schmid and Conrad 1982; Bentz et al. 1983). If they are primary gene products, the short-chain collagens may have emerged prior to the evolution of the full-length interstitial collagen gene. Alternatively, they may be due to extensive deletions in a copy of the full-length ancestral gene as a result of intragenic recombination events. Should the short-chain collagens be synthesized initially as procollagens with amino propeptides homologous to those in the interstitial collagens, the latter explanation would be preferred. Long-chain collagen may have diverged from other members of the collagen gene family by continued sequence amplification within the primordial gene after the full-length ancestral gene had been generated. Long-chain collagen may also have been formed later, by an unequal crossover event.

It is interesting that the α-fetoprotein/albumin genes appear to have been assembled from 27-bp and 54-bp repeating units (Alexander et al. 1984), and that a salivary gland protein gene in *Chironomus* is composed of 9-bp repeating units (Case et al. 1983). Perhaps these genes have organizations similar to that of the collagen gene because there are some structural constraints in eucaryotes on the size of repeats that can be amplified.

Acknowledgments. We wish to thank L. Chao, D. Rosenthal, R. Barstead, and H. Hoyle for helpful discussions. Part of this work was supported by a grant from the Shriners Hospital for Crippled Children to J.L.D.

References

Alexander F, Young PR, Tilghman SJ (1984) Evolution of the albumin: α-fetoprotein ancestral gene from the amplification of a 27 nucleotide sequence. J Mol Biol 173:159–176

Baltimore D (1981) Gene conversion: some implications for immunoglobin genes. Cell 24:592–594

Bentz H, Morris NP, Murray LW, Sakai LY, Hollister D, Burgeson RE (1983) Isolation and partial characterization of a new human collagen with an extended triple-helical structural domain. Proc Natl Acad Sci USA 80:3168–3172

Bernard MP, Chu M-L, Myers JC, Ramirez F, Eikenberry EF, Prockop D (1983) Nucleotide sequences of cDNAs for the pro α1 chain of human type 1 procollagen. Statistical evaluation of structures which are conserved during evolution. Biochemistry 22:5213–5223

Bornstein P, Sage H (1980) Structurally distinct collagen types. Annu Rev Biochem 49:957–1003

Case ST, Summers RL, Jones AG (1983) A variant tandemly repeated nucleotide sequence in Balbiani ring 2 of *Chironomus tentans*. Cell 33:555–562

Chu M-L, Williams CJ, Pepe G, Hirsch JL, Prockop DJ, Ramirez F (1983) Internal deletion in a collagen gene in a perinatal lethal form of osteogenesis imperfecta. Nature 304:78–80

Chu M-L, de Wet W, Bernard M, Ding J-F, Morabito M, Myers J, Williams C, Ramirez F (1984) Human pro α1(I) collagen gene structure reveals evolutionary conservation of a pattern of introns and exons. Nature 310:337–340

Doering JL, Jelachich ML, Hanlon KM (1982) Identification and genomic organization of human $tRNA^{Lys}$ genes. FEBS Lett 146:47–51

Eyre DR (1980) Collagen molecular diversity in the body's protein scaffold. Science 207:1315–1322

Fienberg SE (1977) The analysis of cross-classified categorial data. MIT Press, Cambridge

Fuller F, Boedtker H (1981) Sequence determination and analysis of the 3′ region of chick pro-α1(I) and pro-α2(I) collagen messenger ribonucleic acids including the carboxy-terminal propeptide sequences. Biochemistry 20:996–1006

Hofmann H, Fietzek PP, Kuhn K (1978) The role of polar and hydrophobic interactions for the molecular packing of type I collagen: a three-dimensional evaluation of the amino acid sequence. J Mol Biol 125:137–165

Hofmann H, Fietzek PP, Kuhn K (1980) Comparative analysis of the sequences of the three collagen chains α1(I), α2(I), and α1(III). Functional and genetic aspects. J Mol Biol 141:293–314

Kramer JM, Cox GN, Hirsh D (1982) Comparisons of the complete sequences of two collagen genes in *Caenorhabditis elegans*. Cell 30:599–606

Manning RF, Gage LP (1980) Internal structure of the silk fibroin gene of *Bombyx mori*. II. Remarkable polymorphism of the organization of crystalline and amorphous coding sequences. J Biol Chem 255:9451–9457

McLachlan AD (1976) Evidence for gene duplication in collagen. J Mol Biol 107:159–174

Monson JM, McCarthy BJ (1981) Identification of a Balb/c mouse pro α1(I) procollagen gene: evidence for insertions or deletions in gene coding sequences. DNA 1:59–69

Monson JM, Natzle JE, Friedman J, McCarthy BJ (1982) Expression and novel structure of a collagen gene in *Drosophila*. Proc Natl Acad Sci USA 79:1761–1765

Naora H, Deacon NJ (1982) Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes. Proc Natl Acad Sci USA 79:6196–6200

Peterson RC, Doering JL, Brown DD (1980) Characterization of two *Xenopus* somatic 5S DNAs and one minor oocyte-specific 5S DNA. Cell 20:131–141

Piez KA, Torchia DA (1975) Possible contribution of ionic clustering to molecular packing of collagen. Nature 258:87

Schmid TM, Conrad HE (1982) A unique low molecular weight collagen secreted by cultured chick embryo chondrocytes. J Biol Chem 257:12444–12450

Slightom JL, Blechl AE, Smithies O (1980) Human fetal $^{G}\gamma$ and $^{A}\gamma$ globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21:627–638

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517

Tate V, Finer H, Boedtker H, Doty P (1983) Procollagen genes: further sequence studies and interspecies comparisons. Cold Spring Harbor Symp Quant Biol 47:1039–1049

Trus BL, Piez KA (1976) Molecular packing of collagen: three-dimensional analysis of electrostatic interactions. J Mol Biol 108:705–732

Wozney J, Hanahan D, Tate V, Boedtker H, Doty P (1981) Structure of the pro α2(I) collagen gene. Nature 294:129–135

Yamada Y, Avvedimento VE, Mudryj M, Ohkubo H, Vogeli G, Irani M, Pastan I, deCrombrugghe B (1980) The collagen gene: evidence of its evolutionary assembly by amplification of a DNA segment containing an exon of 54 bp. Cell 22:887–892

Yamada Y, Liau G, Mudryj M, Obici S, deCrombrugghe B (1984) Conservation of the sizes for one but not another class of exons in two chick collagen genes. Nature 310:333–337

Received September 28, 1984/Revised July 6, 1985