Nucleotide Sequences of Complementary Deoxyribonucleic Acids for the Proα1 Chain of Human Type I Procollagen. Statistical Evaluation of Structures That Are Conserved during Evolution†

Michael P. Bernard, Mon-Li Chu, Jeanne C. Myers, Francesco Ramirez, Eric F. Eikenberry, and Darwin J. Prockop*

ABSTRACT: Nucleotide sequences were determined for two cloned cDNAs encoding for over three-fourths of the proα1(I) chain of type I procollagen from man. Comparison with previously published data on amino acid sequences of the α1(I) chain of type I collagen made it possible to examine mutations in the transcribed products of the gene which have occurred during the evolution of man, calf, rat, mouse, and chick. Comparison of the nucleotide sequences with the corresponding sequences of cDNAs from chick [Fuller, F., & Boedtker, H. (1981) *Biochemistry 20*, 996] and with cDNAs for the proα2(I) chain from man [Bernard, M. P., Myers, J. C., Chu, M.-L., Ramirez, F., Eikenberry, E. F., & Prockop, D. J. (1983) *Biochemistry 22*, 1139] demonstrated that selective pressure during evolution for 250 million or more years acted more strongly on the structure of the proα1(I) chain than on the proα2(I) chain. To improve the reliability of the comparison, the nucleotide sequences were examined with a modification of previous procedures for evaluating mutations in replacement sites and silent sites. The corrected divergence for replacement sites between the α1(I) chains was 6 ± 0.8% whereas it was 15 ± 1.9% for the α2(I) chains. The C-propeptide domain of the proα1(I) chain was also highly conserved with a corrected divergence at replacement sites of 5 ± 0.9%, a value that was not distinguishable from the value previously found for the C-propeptide of the proα2(I) chain. Therefore, a large part of the structure of both C-propeptides appears to be under selective pressure. Inspection of changes in the C-propeptide of the proα1(I) chain suggested that there was a highly conserved region around the carbohydrate attachment site similar to the highly conserved region of 37 amino acids previously found in the C-propeptide of the proα2(I) chain. Two statistical tests, however, were unable to confirm nonrandom distribution of changes in the C-propeptide of the proα1(I) chain. The same tests established the presence of a nonrandom distribution in nucleotide changes of the C-propeptide of the proα2(I) chain. The 3′-noncoding region of the cDNA for proα1(I) of human type I procollagen showed no homology with the same region in the chick. Analysis of codon usage for the α1(I) chain indicated the same third base preference for U and C in codons for Gly, Pro, and Ala previously noted for the chick α1(I) chain. The data on the corrected divergence at replacement sites for the proα1(I) and proα2(I) genes in man, mouse, and chick were used to estimate the time since the divergence of the proα1(I) and the proα2(I) genes. The results suggested that the divergence occurred 950 ± 120 million years ago. Since this date precedes several estimates for the first appearance of metazoa, it is possible that the proα genes duplicated before the first multicellular organisms arose. Alternatively, the assumptions for estimating the time of gene duplication based on the evolutionary clock hypothesis may not be valid for collagen genes.

A large part of the primary structure of the α1 and α2 chains of type I collagen was defined by Edman degradation of peptide fragments of the protein obtained from several species [see Kang & Gross (1970), Fietzek & Kühn (1973), Fietzek et al. (1974a,b), Highberger et al. (1975, 1982), Dixit el al. (1977a,b, 1979), Hofmann et al. (1978), Bornstein & Traub (1979), and Piez (1976, 1980)]. More recently, nucleotide sequencing of cloned DNAs for type I procollagen, the precursor of type I collagen, has provided further information about the primary structure of the protein and the coding sequences of the gene (Fuller & Boedtker, 1981; Yamamoto et al., 1980; Showalter et al., 1980; Wozney et al., 1981; Dickson et al., 1981; Olsen & Dickson, 1981). We have reported the nucleotide sequences of cloned cDNAs coding for more than half of the proα2 chain of type I procollagen from man (Myers et al., 1981; Bernard et al., 1983). Comparison of these data with previously published data on homologous cDNAs from chick embryos (Fuller & Boedtker, 1981) made it possible to examine evolution of the gene in two species which have diverged for 250–350 million years. The results demonstrated that the amino acid sequences of the α2(I) chain domain are not as highly conserved as those of the C-propeptide domain of the same chain. Around the carbohydrate attachment site in the central portion of the C-propeptide, there was complete conservation of 50 out of 51 amino acids. In addition, comparison of the human and chick cDNAs revealed several classes of conservation of nucleotide sequences which had no apparent effect on the primary structure of the protein. The conservation of nucleotide sequences inlcuded a preference for U in the third base position of codons for glycine, proline, and alanine, a high degree of nucleotide conservation in the 51 amino acid highly conserved region of the C-propeptide, and a moderate degree of conservation of the 3′-noncoding region of the mRNA.

Here we report nucleotide sequences of two cloned cDNAs coding for two-thirds of the proα1 chain of type I procollagen from man and part of the 3′-noncoding region of the mRNA (Chu et al., 1982). Comparison with previously published data on amino acid sequences of α1(I) chains from several species and with nucleotide sequences on cloned DNAs for the same chain from chick and mouse made it possible to define further

† From the Departments of Biochemistry and Obstetrics/Gynecology, University of Medicine and Dentistry of New Jersey-Rutgers Medical School, Piscataway, New Jersey 08854. *Received May 20, 1983.* This work was supported in part by National Institutes of Health Grant AM 16516 and grants from The March of Dimes-Birth Defects Foundation and the Lalor Foundation.

0006-2960/83/0422-5213$01.50/0

features of the type I procollagen genes that are conserved through evolution.

Materials and Methods

Enzymes and Other Materials. Restriction endonculeases were purchased from New England Biolabs and Bethesda Research Laboratories. T4 polynucleotide kinase was purchased from Bethesda Research Laboratories. Labeled nucleotides were purchased from Amersham Corp.

DNA Sequence Determination. DNA sequencing was carried out essentially as described by Maxam & Gilbert (1980). The 5′ ends of restriction fragments were labeled with [γ-^{32}P]ATP and T4 polynucleotide kinase. Either the labeled fragments were cleaved with a second restriction endonuclease or the strands were separated by electrophoresis. The polyacrylamide gels for DNA sequencing were 0.4 mm thick by 40 cm long and were run at 1100–1600 V. One 8% gel and one or two 5% gels were used for each determination. To determine the first few nucleotides in a fragment, a 20% gel 0.8 mm thick was employed.

Experiments involving recombinant DNA were performed in P1-EK1 containment in accordance with guidelines of the National Institutes of Health.

Calculation of the Corrected Divergence and Its Sampling Error. The divergence of nucleotide sequences was examined by a modification of the method of Perler et al. (1980) in which the reliability of the statistical comparisons was improved by replacing the arbitrarily weighted average divergence by the variance-weighted average (Clifford, 1973). Each nucleotide was first evaluated in terms of whether a change in the base present at a given site would produce replacement of an amino acid or be silent in its effect on the protein. Both replacement and silent sites were divided into three classes according to the number of different ways (1, 2, or 3) in which a particular replacement or silent mutation could be realized. The aligned nucleotide sequences were then compared to determine the number of silent and replacement mutations that had occurred in each of the three classes. The observed fraction of nucleotide sites of class i that differ was expressed as λ_i, $i = 1, 2, 3$. The values were corrected for the average probability that more than one mutation had occurred at each site to give K_i, the corrected number of substitutions per site of class i, according to

$$K_i = -\frac{a_i}{b_i} \ln (1 - b_i\lambda_i) \qquad i = 1, 2, 3$$

where $a_1 = 3$, $a_2 = 1.5$, $a_3 = 1$, $b_1 = 2$, $b_2 = 1.5$, and $b_3 = {}^4/_3$ (Perler et al., 1980) for both silent and replacement sites. Finally, the K_i's were combined to form a variance-weighted average corrected divergence for silent and for replacement mutations according to

$$K = \sum_{i=1}^{3} w_i K_i / \sum_{i=1}^{3} w_i$$

where the weights, w_i, are the reciprocals of the sampling variance for each class of mutation (see below).

To assess the statistical significance of differences in corrected divergence, the sampling variance was calculated by an extension of the method of Kimura & Ohta (1972) which incorporated the distinction among the three classes of mutations. The sampling variance of K_i is given by

$$\sigma_i^2 = \frac{a_i^2\lambda_i(1 - \lambda_i)}{(1 - b_i\lambda_i)^2 L_i} \equiv \frac{1}{w_i}$$

where L_i is the number of sites of class i in the sequence. The variance of K, the corrected divergence averaged over the three classes, is given by

$$\sigma^2 = (\sum_{i=1}^{3} w_i)^{-1}$$

The sampling error reported in Table IV is σ, the square root of this variance.

The variance-weighted average provided a minimum variance estimate of the overall corrected divergence between two sequences. The use of this average reduced, in some cases by an order of magnitude, the sampling error of the divergence by comparison to the previously used weighted average (Perler et al., 1980). This also caused small reductions in the estimates of replacement site and silent site divergence previously reported for the proα2(I) chain (Bernard et al., 1983). It may be noted that the maximal value for the corrected divergence is indefinitely large, corresponding to an indefinitely large number of substitution events. However, in these cases the sampling variance is also indefinitely large, and such occurrences have a very small weight in the overall divergence. Hence, with the formulation used here, it was unnecessary to exclude arbitrarily the rare classes of mutations from the calculations as was previously done (Perler et al., 1980; Bernard et al., 1983).

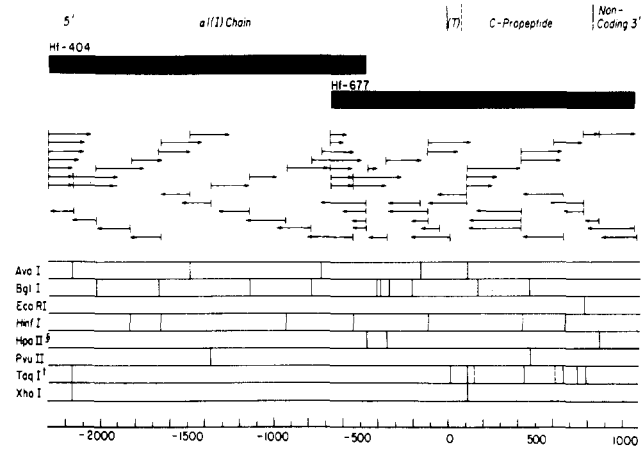

FIGURE 1: Restriction map and sequencing strategy for two cloned cDNAs for the proα1(I) chain of normal human type I procollagen. Horizontal lines and arrows indicate fragments sequenced. (T) C-terminal telopeptide; (§) partial restriction map; (†) dotted lines indicate restriction site without cleavage. Nucleotides (scale at bottom) are numbered as suggested by Fuller & Boedtker (1981) with the last nucleotide coding for the last residue of the α-chain domain designated "−1". As noted in text, the 5′ end of Hf-404 encodes amino acid residue 247 of the α1(I) chain.

Results

Sequencing Strategy. Nucleotide sequences were obtained from two cloned cDNAs prepared from normal human skin fibroblasts (Chu et al., 1982). Each of the clones was about 1.8 kilobases in length and the nucleotide sequences in the two clones overlapped by 0.22 kilobase (Figure 1). A restriction endonuclease map was generated and used to develop a strategy for nucleotide sequencing. Essentially all the sequences were confirmed by sequencing the two strands of the DNA. As noted in Figure 1, one of the clones (Hf-677) included part of the 3′-noncoding region of the mRNA. The 5′ end of the other clone (Hf-404) coded for amino acid residue 247 of the α1(I) chain. Together the two cDNAs encoded for 768 amino acid residues or 76% of the α1(I) chain, all 26 amino acid residues of the C-terminal telopeptide, all 246 amino acid residues of the C-propeptide, and 224 base pairs of the 3′-noncoding region. Restriction maps of several smaller,

Table I: Common Amino Acid Substitutions between Human and Chick α1(I) Chains

human amino acid	chick amino acid[a]			
	Ala	Ser	Pro	Val
Ala		0	7	2
Ser	8		1	0
Pro	9	0		0
Val	2	0	2[b]	

[a] Changes indicated account for 31 out of 54 amino acid substitutions. As noted in text, 28 of the substitutions involved a change of Ala in one sequence or the other. [b] Change in amino acid requires mutation of two nucleotides; the other changes can be explained by mutation of a single nucleotide only.

overlapping cDNAs obtained from the same cloning experiments were consistent with the restriction map shown in Figure 1. Therefore, there was no evidence of any nucleotide rearrangement during the preparation of the cloned cDNAs as was previously encountered in preparing cDNAs of chick proα1(I) and proα2(I) chains [see Fuller & Boedtker (1981)].

Comparison of the Amino Acid Sequence of the Human α1(I) Chain with Amino Acid Sequences from Calf, Rat, Mouse, and Chick. The amino acid sequences encoded for by the cDNAs for the α1(I) chain were compared with published data for the same structure from four other species (Figures 2 and 3). As expected, there was complete conservation of Gly in the first position of the repeating tripeptide sequence Gly-X-Y which comprises the triple-helical or α domain of collagen. As also expected, there were essentially no changes in distribution of charged residues. At two positions (amino acid residues 384 and 488) there was a substitution of Asp for Asn, but this apparent discrepancy may be explained by the difficulty of distinguishing these two amino acids by analysis of peptide fragments. At two additional sites (amino acid residues 543 and 548) there was a substitution of Gln for Glu which probably has a similar explanation. At one site (residue 677) there was a substitution of Asn for Glx, and at another site (residue 995) there was substitution of Asp for Glu. There were reasonably extensive substitutions of other amino acids, but most of these were highly conservative. Of 54 amino acid substitutions between the human and chick α1(I) chains, 31 involved interconversions of Ala, Ser, Pro, and Val (Table I). Nineteen of the amino acid differences involved a change in Ala in the chick sequence to Ser, Pro, or Val in the human, and another nine of the differences involved a change of Ala in the human sequence to Pro or Val in chick.

As predicted from analysis of cyanogen bromide fragments [see Bornstein & Traub (1979) and Piez (1976)], the Met found in position 929 of the chick α1(I) chain was not found in the human chain (Figure 3). The region containing the cleavage site for vertebrate collagenase (between positions 775 and 776) was highly conserved, but there were several other regions of the α1(I) chain that showed similar degrees of conservation.

There were three amino acid substitutions between calf and human protein in the telopeptides, and these substitutions were conservative in that they involved a substitution of Ala for Gly, Phe for Tyr, and Phe for Leu.

Comparison of Amino Acid and Nucleotide Sequences of the α Domain and C-Propeptide Domain between Man and Chick. Over half of the nucleotide sequences obtained here overlapped with nucleotide sequences determined for the chick proα1(I) chain (Fuller & Boedtker, 1981). Therefore, the overlapping nucleotide sequences in the two species were

Table II: Comparison of Amino Acid and Nucleotide Differences between Human and Chick Proα1(I)

	amino acid replacements	nucleotide changes (uncorrected values)	
		replacement sites	silent sites
α1(I) chain[a]	59/794[c]	16/498	103/229
α1(I) chain *less* Gly[b]	58/512[c]	15/298	58/134
telopeptide	1/26		
C-propeptide	24/246	31/627	72/246

[a] Includes the C-terminal telopeptide. [b] Does not include glycine in every third position or the C-terminal telopeptide. [c] Comparison based on amino acid sequences determined by analysis of peptide fragments covering amino acid residues 247-1040. Other data are based on nucleotide sequences only and begin with amino acid residue 814 of the α1(I) chain. Denominator indicates total number of amino acids or sites.

Table III: Comparison of Amino Acid and Nucleotide Differences between Human and Chick Proα1(I) and Proα2(I) Chains

	amino acid replacements (% difference)[a]		nucleotide changes (% corrected divergence)[b]			
			replacement sites		silent sites	
	α1	α2	α1	α2	α1	α2
α-chain domain	7	22	6 ± 0.8	15 ± 1.9	81 ± 11	51 ± 7
α-chain domain *less* Gly	10	32	5 ± 1.4	23 ± 3.1	79 ± 14	47 ± 9
C-propeptide	10	11[b]	5 ± 0.9	8 ± 1.3[c]	62 ± 9	66 ± 9[c]
3′ noncoding					ind.[d]	35 ± 5[e]
β-globin[f]	31		23		70	
insulin[f]	14		8		122	

[a] Amino acid replacement values for α1 chains are based on amino acid sequences determined by analysis of peptide fragments covering amino acid residues 247-1040 (Figures 2 and 3). All other data are based on nucleotide sequences only and begin with amino acid residue 814 of the α1(I) and α2(I) chains (Figure 4; Bernard et al., 1983). [b] Values are variance-weighted means ± sampling error (see Materials and Methods). [c] Value calculated after optimal alignment of sequences. To obtain optimal alignment, insertion of four amino acids and deletion of two amino acids in human C-propeptide are ignored (see Figure 3; Bernard et al., 1983). [d] Apparent divergence is indeterminately large. [e] Value for homologous sequences after alignment (223 of 262 bases) (Bernard et al., 1983). [f] Values from Efstratiadis et al. (1980). Values will differ slightly if recalculated as variance-weighted means with the procedure introduced here.

compared (Figure 4). For the analysis, a modification of the procedure of Perler et al. (1980) was used, and the data were expressed as corrected divergence for mutations in replacement sites and in silent sites. In addition, we developed an extension of the method of Kimura & Ohta (1972) to estimate the sampling error (see Materials and Methods).

As indicated in Table II and in Figure 4, there were 16 nucleotide changes in replacement sites in nucleotides of the α domain of the proα1(I) chain and 103 mutations in silent sites in the nucleotides of the same domain. There were 31 mutations in replacement sites of the C-propeptide and 72 mutations in silent sites in the same domain. The values for divergence of the α1 domain demonstrated greater conservation of amino acid sequences than was found previously in the α2 domain of the proα2(I) chain from the same two species (Table III). When expressed either as the percent replacement in amino acids or the percent corrected divergence at replacement sites, the values for the α1(I) chain were less than half of the values obtained for the α2(I) chain. Calculation

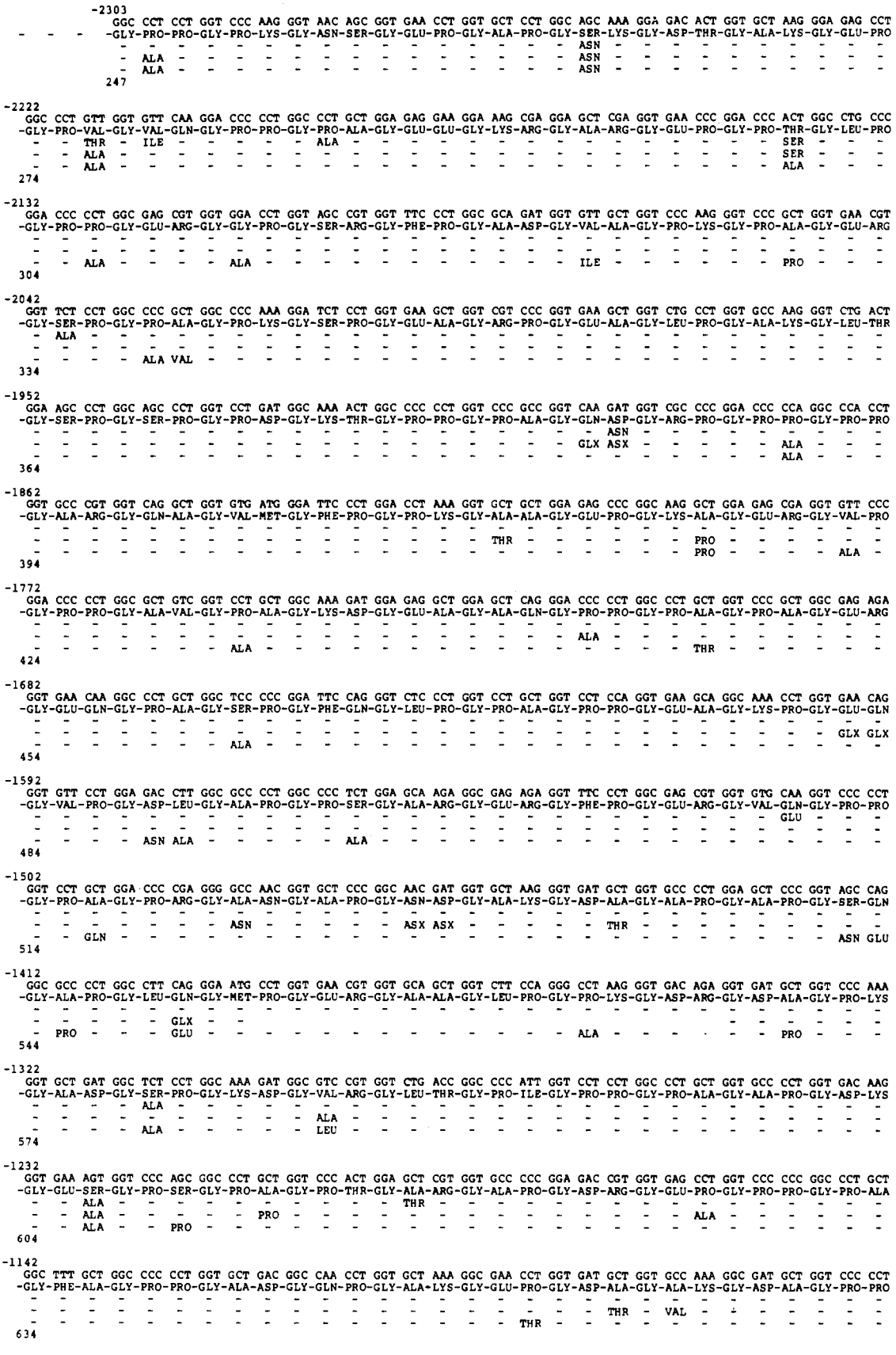

FIGURE 2: Comparison of the amino acid sequences for $\alpha 1$(I) chains among man, calf, rat or mouse, and chick. First line: nucleotide sequence of cDNAs for human pro$\alpha 1$(I). Second line: amino acid sequence encoded by cDNAs for human pro$\alpha 1$(I). Third line: amino acid sequence of calf $\alpha 1$(I) (Hofmann et al., 1978). Fourth line: amino acid sequence of $\alpha 1$(I) for rat from amino acid residue 247 to the Met residue at 551 (Piez, 1976) and of $\alpha 1$(I) for mouse from amino acid residue 568 to 963 (Monson & McCarthy, 1982). Fifth line: amino acid sequence for the chick $\alpha 1$(I) chian (Fuller & Boedtker, 1981; Highberger et al., 1982). *Dashes* represent identity with human amino acid sequence.

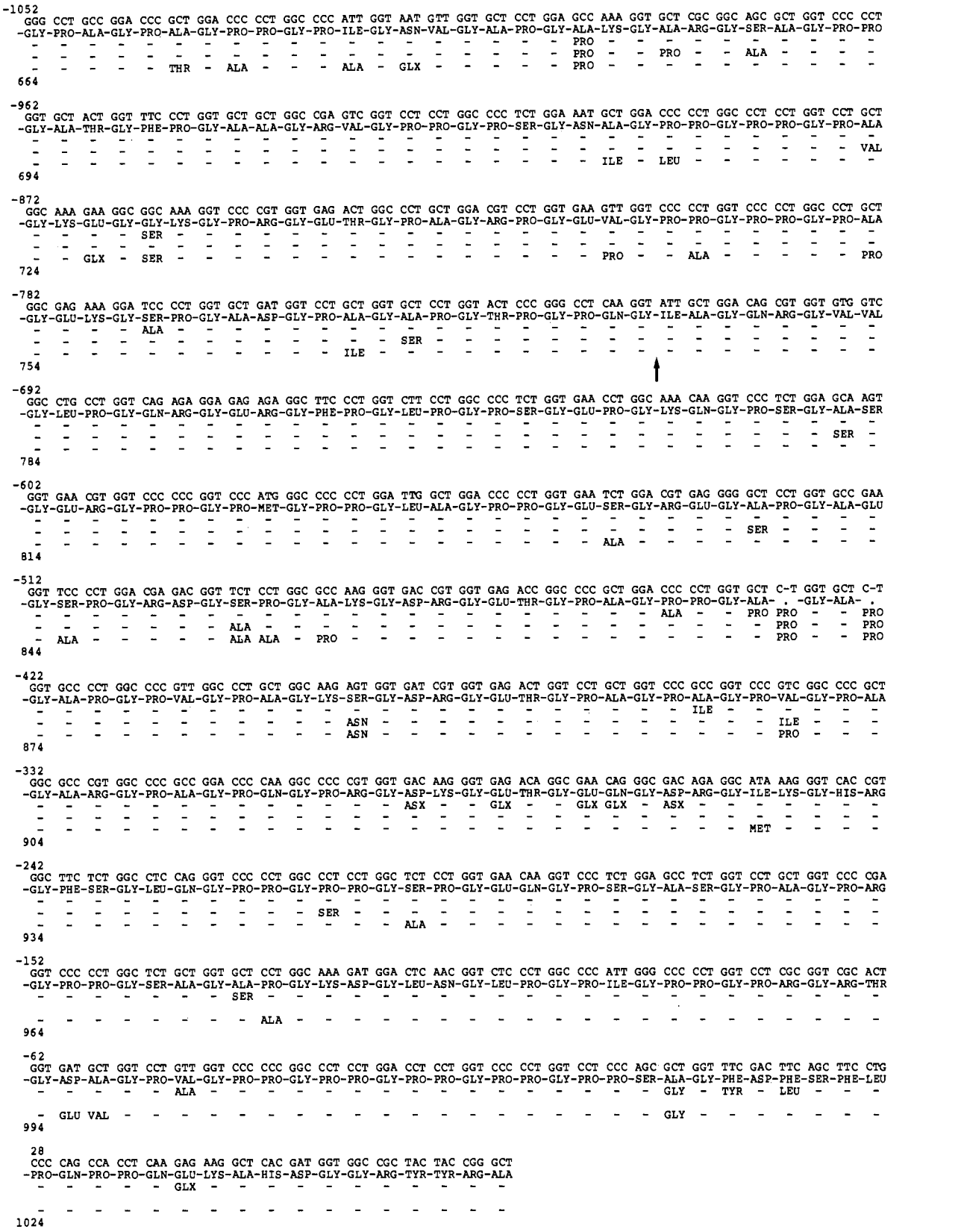

FIGURE 3: Comparison of the amino acid sequences for α1(I) chains among man, calf, rat or mouse, and chick (continued from Figure 2). Lines 1–5: as in Figure 2. Vertical arrow: cleavage site for mammalian collagenase between amino acid residues 775 and 776. Dashes at nucleotides −433 and −424: Bases not clearly defined by nucleotide sequencing.

of the sampling error demonstrated that the difference in corrected divergence between the α1(I) and α2(I) chains for replacement sites was significant. The corrected divergence for silent sites in the α1(I) chains, however, was the same as the values previously obtained for the α2(I) chain. Evaluation of the C-propeptides indicated that the corrected divergence for replacement sites was indistinguishable for the proα1 and proα2 chains. Again, the values for corrected divergence in silent sites were the same for the C-propeptides of the two chains.

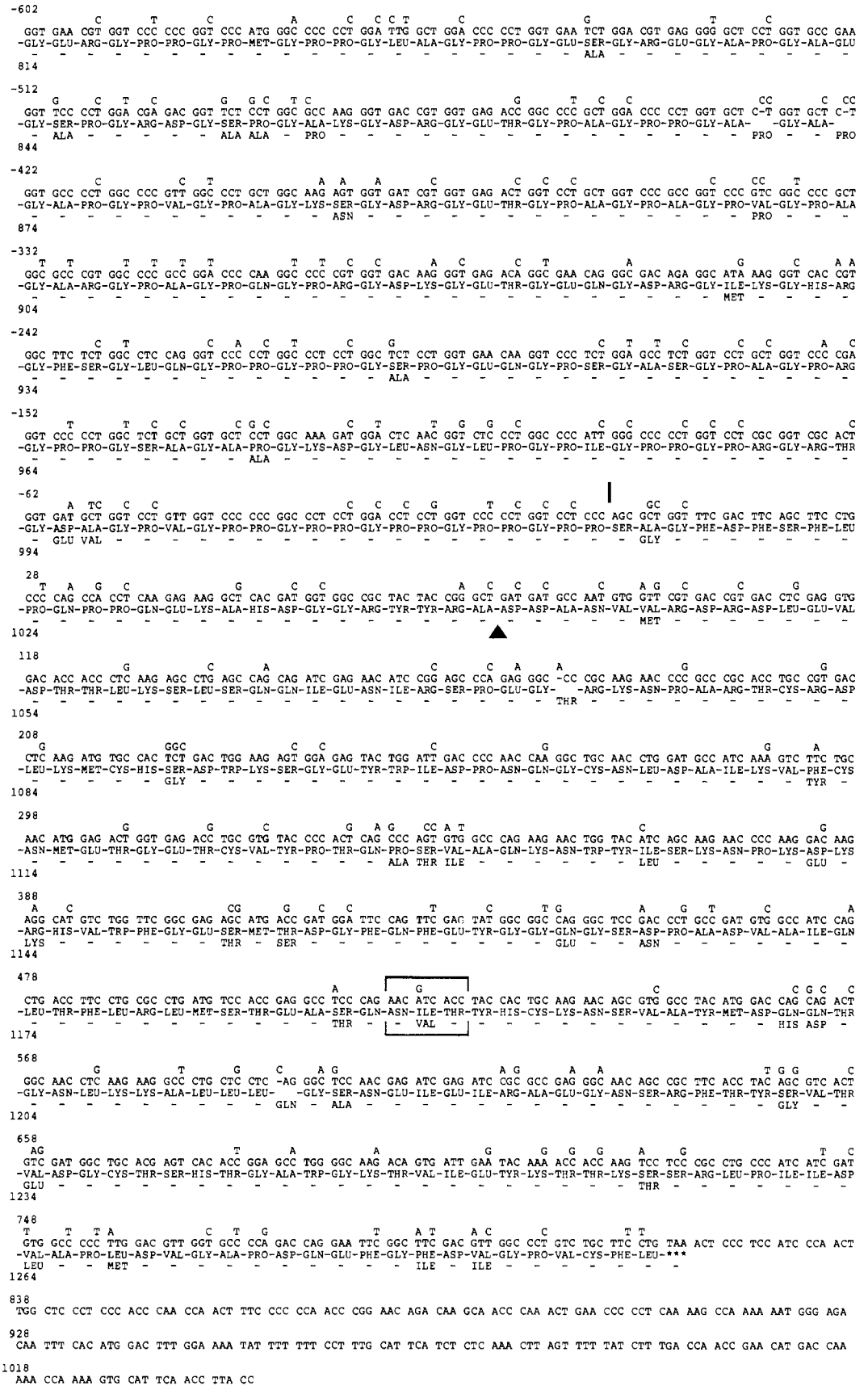

FIGURE 4: Comparison of human and chick cDNAs for proα1(I). First line: nucleotide sequence in the chick cDNAs where this differs from the human. Second line: nucleotide sequence obtained from the human cDNAs for proα1(I). Third line: amino acids encoded by the human cDNAs. Fourth line: Amino acid sequence for the chick proα1(I) chain where this differs from the human. Dashes at nucleotides −433, −424, 175, and 595: bases not clearly defined by nucleotide sequencing. (Vertical bar) End of α-chain domains between amino acid residues 1014 and 1015. (▲) Site at which C-propeptide is cleaved during the conversion of procollagen to collagen. (Asterisks) Termination codon for translation. (□) Carbohydrate attachment site.

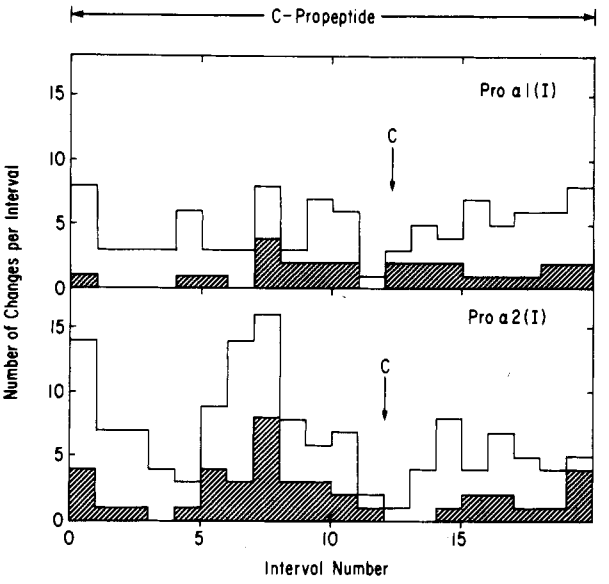

FIGURE 5: Distribution of differences in amino acids and nucleotides between human and chick C-propeptides. The C-propeptides were arbitrarily divided into intervals of 12 amino acids (36 nucleotides). Interval 0–1 for the proα1(I) C-propeptide consists of amino acid residues 1041–1052 in Figure 4 etc. Because they did not constitute a complete interval, the last six amino acids in the proα1(I) C-propeptide and the last three amino acids in the proα2(I) C-propeptide were ignored. The values for number of changes include the insertion of four amino acids and the deletion of two amino acids in the human proα2(I) C-propeptide (Bernard et al., 1983). (C) carbohydrate attachment site; (open areas) nucleotide changes; (shaded areas) amino acid changes.

Comparison of the 3′-noncoding regions of the proα1(I) chains of man and chick indicated that there was no homology in the sequences. In contrast, comparison of the 3′-noncoding region for the proα2(I) genes indicated (Bernard et al., 1983) a corrected divergence of 35 ± 5% after the alignment of the sequences (Table III).

Search for a Highly Conserved Region in the C-Propeptide. Comparison of the amino acid and nucleotide sequences of the human and chick C-propeptides of the proα2(I) chain indicated that there was a highly conserved region around the carbohydrate attachment site (Bernard et al., 1983). Therefore, we examined the proα1(I) C-propeptide for the same type of conservation. Although the structure around the carbohydrate attachment site of the human proα1(I) chain was homologous with the chick, the presence of a highly conserved region was less apparent than in the proα2 C-propeptide (Figure 5).

For a more precise evaluation of conservation, two kinds of statistical tests were carried out at both the amino acid and nucleotide levels. In the first, the data in Figure 5 were analyzed with a χ^2 test against the null hypothesis that each arbitrary interval of 12 amino acids (36 nucleotides) contained the same average number of changes. For the C-propeptide of the proα1(I) chain, the χ^2 values indicated that the distribution of changes was not distinguishable from a uniform distribution. For the C-propeptide of the proα2(I) chain, however, χ^2 was 27.3 for amino acid mutations and 31.6 for nucleotide changes which, with 18 deg of freedom, gave P values of <0.06 and <0.05, respectively.

The second test was a χ^2 test of the observed distribution of distances between mutations against the expected values derived from the Poisson interval distribution (Evans, 1955). For the C-propeptide of the proα1(I) chain, the χ^2 test indicated that the observed distribution of mutations at either the amino acid level (not shown) or nucleotide level (Table IV) could not be distinguished from a random distribution. For the C-propeptide of proα2(I) chain, the observed distribution of distances between nucleotide mutations was highly significantly different from a random distribution. However, the distribution of distances among amino acid changes was not (not shown).

Table IV: Comparison of Observed Distances between Changes in Human vs. Chick C-Propeptide with Expected Values Derived from the Poisson Interval Distribution

distance between changes	proα1 C-propeptide		proα2 C-propeptide	
	observed no.	expected no.[b]	observed no.	expected no.[b,c]
1	14	19.1	19	25.5
2	8	10.7	14	13.8
3	16	9.3	26	11.8
4–5	12	15.2	10	18.5
6–8	17	16.2	16	18.6
9–13	19	15.7	23	16.5
14–20	12	9.8	6	9.0
21–23	3	5.0	3	3.7
34–	1	1.0	1	0.5
av distance	7.25		6.17	
χ^2	9.6		27.2	
p	~0.25		<0.001	

[a] Intervals were selected to obtain relatively uniform distribution among the classes. [b] Evans (1955). [c] Values ignore insertion of four amino acids and deletion of two amino acids found in comparison of human C-propeptide with chick C-propeptide of the proα2(I) chain.

Table V: Comparison of Amino Acid and Nucleotide Differences between Human and Mouse Proα1(I) Chains

	amino acid replacements (% difference)	nucleotide changes (% corrected divergence)[b]	
		replacement sites	silent sites
α1(I) domain	4.3	2.3 ± 0.5	33 ± 3.7
α1(I) *less* Gly	6.5	3.3 ± 0.7	33 ± 4.7
β-globin[a]		13	49

[a] Values from Efstratiadis et al. (1980) for human β-globin gene and the mouse β^{maj} gene. [b] Values calculated as in Table III. Data for mouse proα1(I) chain are from Monson et al. (1982).

Comparison of Amino Acid and Nucleotide Differences between Man and Mouse Proα1(I). Recently, nucleotide sequences were reported for several exons of the mouse proα1(I) gene (Monson et al., 1982). Therefore, we carried out a comparison between the mouse and human nucleotide sequences (Figure 6 and Table V). The values for amino acid replacements and for corrected divergence for replacement and silent sites were much less than comparable values for the same gene in man and chick, an observation consistent with the smaller evolutionary separation of man and mouse.

Codon Usage. Analysis of the codon usage for the α1(I) chain (Table VI) indicated the same third base preference for U and C in codons for Gly, Pro, and Ala previously noted for chick α1(I) chain (Fuller & Boedtker, 1981). In these codons, there generally was a preference for U over C. In the chick α1(I) chain, however, C was frequently used for the third base in codons for Pro, Ala, and Arg. The data also indicated a strong preference for the codon CGU for Arg in both the α1 and α2 chains of man and the α1 chain of mouse. In the case of the chick α1 chain, there is a preference for the codon CGC for Arg, but the number of Arg codons examined in the chick α1 chain was relatively small.

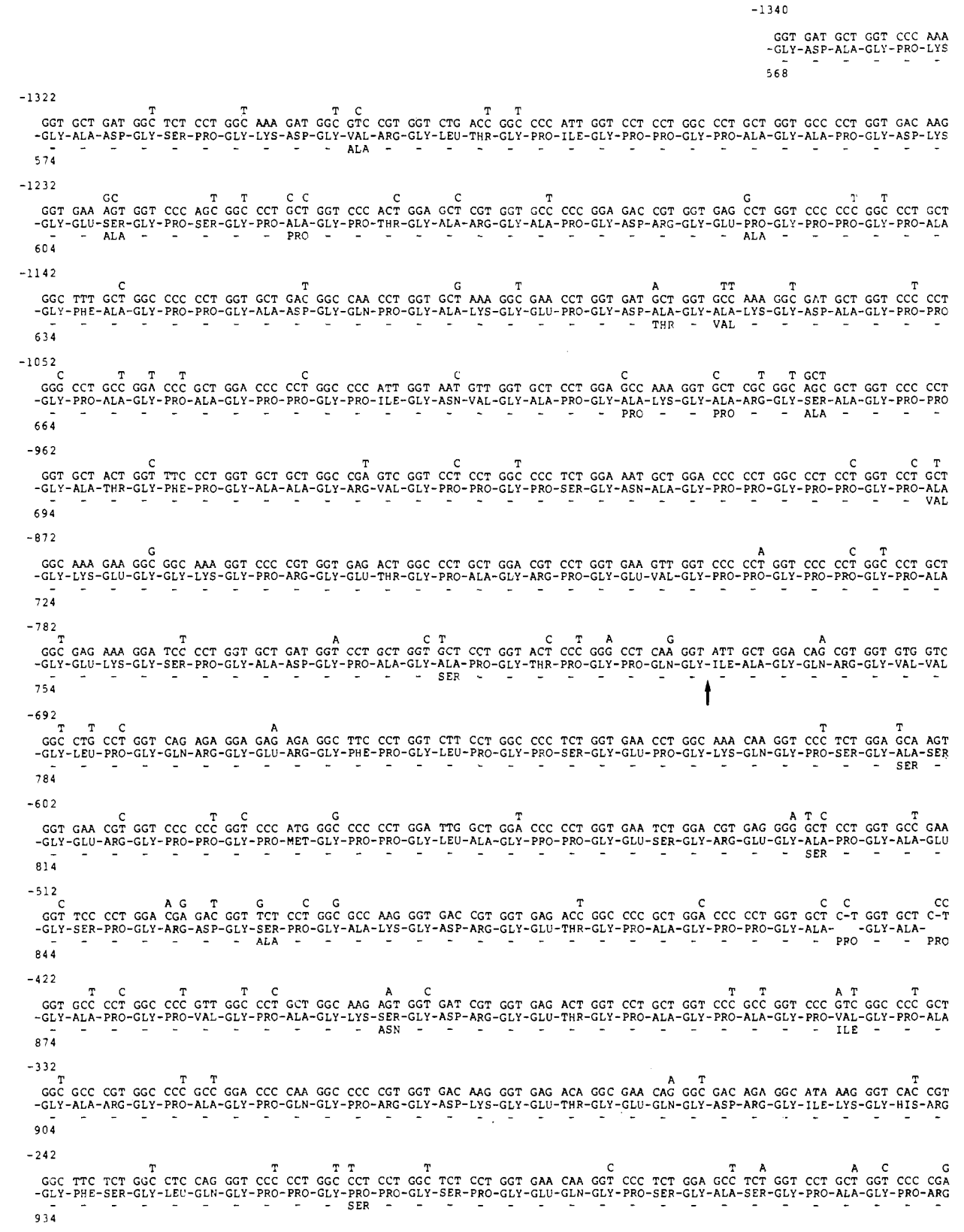

FIGURE 6: Comparison of exons from mouse genomic DNA for proα1(I) with corresponding regions of human cDNA. Mouse nucleotide sequences are from Monson et al. (1982). Vertical arrow: Cleavage site for vertebrate collagenase.

Use of the Evolutionary Clock Hypothesis To Estimate the Time Since the Divergence of the Proα1(I) and Proα2(I) Genes. The values for the corrected divergences of the proα1(I) and proα2(I) genes make it possible to use the procedures recently employed with several other genes (Efstratiadis et al., 1980; Perler et al., 1980) to estimate the time since the two genes diverged. Such estimates are generally based on fossil data, suggesting that the mammalian radiation occurred about 80 million years ago and that the last common ancestor of mammalian and avian species was about 270 million years ago (McKenna, 1975). These two time points were used to calibrate the evolutionary clock for type I procollagen and to estimate the time since the proα1(I) and proα2(I) genes diverged (Figure 7). Three separate estimates

Table VI: Codon Usage in α1(I) and α2(I) Domains[a]

amino acid	α1(I) domain man	α1(I) domain mouse	α1(I) domain chick	α2(I) domain man	α2(I) domain chick	third base or codon
Gly	0.50	0.60	0.57	0.56	0.67	U
	0.30	0.23	0.36	0.20	0.21	C
	0.18	0.14	0.06	0.19	0.12	A
	0.02	0.02	0.01	0.04	0	G
codons examined	261	133	67	166	67	
Pro	0.57	0.56	0.25	0.66	0.75	U
	0.40	0.43	0.69	0.21	0.13	C
	0.03	0.01	0.04	0.10	0.13	A
	0	0	0.02	0.03	0	G
codons examined	179	91	50	92	48	
Ala	0.74	0.80	0.52	0.84	1.00	U
	0.21	0.16	0.48	0.09	0	C
	0.05	0.02	0	0.07	0	A
	0	0.02	0	0	0	G
codons examined	92	44	23	55	14	
Arg	0.50	0.70	0.25	0.33	0.64	CGU
	0.13	0.05	0.58	0.11	0.09	CGC
	0.18	0	0	0.11	0	CGA
	0.03	0.05	0	0.11	0	CGG
	0.18	0.15	0.17	0.22	0.18	AGA
	0	0.05	0	0.11	0.09	AGG
codons examined	40	20	12	27	11	

[a] Data for human α1(I) and α2(I) chains are from Bernard et al. (1983) and Figures 2-4. Data for mouse α1(I) are from Monson et al. (1982). Data for chick α1(I) and α2(I) chains are from Fuller & Boedtker (1981).

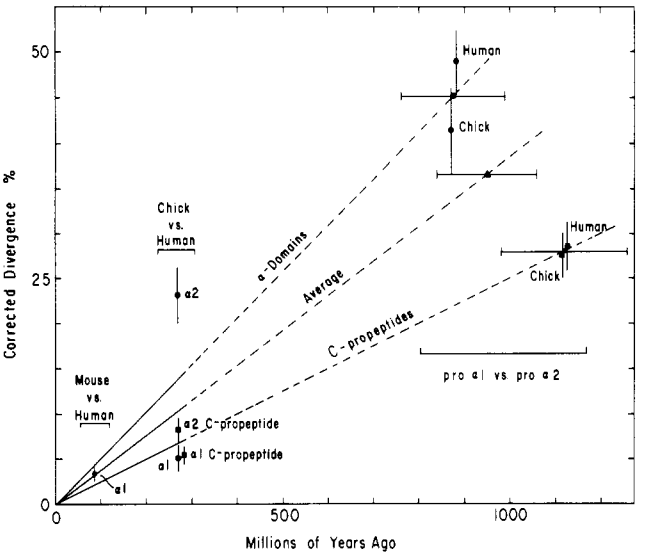

FIGURE 7: Evolutionary clock for type I procollagen. Data for comparison of mouse and human genes at mammalian radiation (80 million years) and at last common ancestor of chick and man (270 million years) are from Tables III and V. Data for corrected divergence between the α1(I)- and α2(I)-chain domains and between the two C-propeptides were calculated with the same procedures. (●) α-chain domains; (■) C-propeptides; (▲) average of values for α-chain domains and C-propeptides. Bars indicate standard error on both ordinate and abscissa. All values were variance-weighted averages, and the uncertainties were propagated onto the estimate (Clifford, 1973).

of the time since the divergence were made: one from the data for the two C-propeptides, a second from the data for the two α-chain domains, and a third from the combined data. The α2 chain clearly diverged at a faster rate than the α1 chain or the C-propeptides. However, there was no clear reason for excluding the data on the α2 chain, and excluding these data did not significantly affect the estimated time. The average value for time since the divergence was 950 ± 120 million years. The other two estimates based on either the α-chain domains or the C-propeptides overlapped this value (Figure 7).

Discussion

The data presented here provide the first extensive information about the primary structure of the proα1(I) chain of human type I procollagen and the nucleotide sequences of the mRNAs for this protein.

The results demonstrate that both the α domain and the C-propeptide of proα1(I) chain have been conserved to a relatively high degree. A high degree of conservation of amino acid sequences in the α-chain domain was previously demonstrated by direct analysis of α chains from mammalian and avian species (Bornstein & Traub, 1979; Hofmann et al., 1978; Mathews, 1980). The data here demonstrate that the amino acid sequence of the C-propeptide of the proα1(I) chain is also highly conserved. The high degree of conservation is even more apparent if the data are expressed as corrected divergence of nucleotides at replacement sites (Perler et al., 1980), a procedure which makes it possible to take into account the number of point mutations required to change a given amino acid and corrects for the probability of multiple mutations. The values of 6 ± 0.8% corrected divergence at replacement sites in the α1(I) chain and 5 ± 0.9% in the C-propeptide of the proα1(I) chain are considerably smaller than the value of 23% determined from a comparison of human and chick β-globin, a protein in which only a small part of the structure is preserved in evolution. The values for the two domains of the proα1(I) chain are in fact less than the 8% divergence between human and chick insulin, a protein which is regarded as being highly conserved. The high degree of nucleotide conservation for the α1(I) chain domain was also seen in the comparison of DNA sequences between man and mouse, species which have diverged for about 80 million years (McKenna, 1975).

Under the assumption that the structural features which are conserved through evolution are required for important biological functions, the results suggest that most of the structure of the proα1(I) chain has a critical biological role. The need to conserve the α domain is consistent with the fact that the chains of collagen are folded into a triple-helical conformation which requires Gly in every third position and a high content of Pro and Hyp. It is also consistent with the fact that correct self-assembly of collagen into fibers requires a precise distribution of charged and hydrophobic amino acids on the surface of the molecule. Since the degree of conservation of the α2(I) chain is less than that of the α1(I), the results suggest that the structure of the α domain of the proα2(I) chain is less critical for the formation of normal fibers from type I collagen than the structure of the α1(I) chain.

The reasons for the relatively high degree of conservation of the C-propeptides of both the proα1 and proα2 chains are less apparent. The C-propeptides are involved in chain selection and chain association during intracellular assembly of procollagen, and they appear to have a role in maintaining the protein in soluble form during transport to the extracellular matrix [see Prockop et al. (1979) and Bornstein & Sage (1980)]. In addition, a variety of other functions have been suggested. The high degree of conservation seen throughout the C-propeptide domains suggests that these biological functions together involve a large part of the structures.

Although the degree of conservation of the C-propeptide

of the proα1 chain was about the same as the degree of conservation seen in the C-propeptide of the proα2 chain, there were several differences. In the C-propeptide of the proα2 chain, the human protein contained an insert of four amino acids near the N-terminus and a deletion of two amino acids in the first third of the protein when compared to the chick C-propeptide (Bernard et al., 1983). No such insertions or deletions were found in the C-propeptide of the proα1 chain. Also, the highly conserved region around the carbohydrate attachment site which was found in the C-propeptide of the proα2 chain was less prominent in the proα1 chain. For this reason, we used statistical procedures to evaluate the amino acid and nucleotide sequences. Two different tests did not demonstrate the presence of a nonrandom distribution of changes in the C-propeptide of the proα1(I) chain, but they did confirm the presence of such a nonrandom distribution in the C-propeptide of the proα2(I) chain. It should be noted that long contiguous sequences of unchanged residues are to be expected on a random basis in homologous sequences of either nucleotides or amino acids. For example, comparison of the human and chick proα2(I) C-propeptides showed that there was a 9% probability of encountering by chance alone an unmutated sequence of 44 nucleotides, the actual length of the most highly conserved region. Thus, a large data base is necessary to identify with certainty a selectively conserved region. It is possible that conservation of the region containing the carbohydrate attachment site of the proα1(I) C-propeptide will become more definitive when the data base can be enlarged to include proα1(I) genes from additional species.

The structure 3′-noncoding region of the message for the proα1(I) chain showed no apparent homology between chick and man. Moderate degrees of homology have been seen in 3′ noncoding of regions of some mRNAs such as those for β-globins (Efstratiadis et al., 1980; Martin et al., 1981), but not of others such as actin (Hanukogu et al., 1983). The lack of any apparent conservation noted here suggests that the 3′-noncoding region of the proα1(I) chain in man may have had a different origin than the same region of the proα1(I) chain in chick. Since the 3′-noncoding region of the proα2(I) mRNA was conserved to a moderate degree (Bernard et al., 1983), it may have a more critical function.

The unusual third base preference for U and C in codons for Gly, Pro, and Ala previously noted in chick cDNAs (Fuller & Boedtker, 1981) is clearly present in the human mRNAs as well as the mouse. The preference for U and C in codons for Gly and Ala was predicted from observations indicating that glycyl-tRNA and alanyl-tRNA recognizing codons with U or C in the third position were preferentially used in tissues synthesizing collagen (Carpousis et al., 1977). There is, however, no apparent explanation for this type of selective codon usage in collagen genes. It may be related to the availability of specific isoaccepting species of tRNAs for translation of procollagen mRNAs or to secondary structural requirements for regulating the translation of the mRNAs.

The data on the corrected divergence at replacement sites for the proα1(I) and proα2(I) chains were used to calibrate an evolutionary clock in the same manner previously used for a number of other genes (Efstratiadis et al., 1980; Perler et al., 1980). The procollagen genes may be ideal for this purpose. The genes are large, they are ancient, and the molecule they code for is probably subject to uniform selective pressure along its whole length because of the lateral interactions required for fiber formation. Therefore, the procollagen genes probably provide a larger and more homogeneous data base for evaluating mutations than most other genes. The data developed here indicated that the proα1(I) and proα2(I) genes diverged 950 ± 120 million years ago. This estimate places the divergence further back in time than the estimates of 560–840 million years ago based on analysis of amino acid sequences of α chains (Bornstein & Traub, 1979; Mathews, 1980). It also places the divergence further in time than the estimate of 500 million years for the divergence of α- and β-globin genes (Efstratiadis et al., 1980). From observations on oxidized iron in sedimentary strata, it was estimated that metazoa appeared about 680 million years ago (Cloud, 1968). Therefore, the estimate here of 950 million years for divergence of the proα genes raises the possibility that the genes duplicated before the first multicellular organisms arose. Alternatively, the data may indicate that the assumptions on which the evolutionary clock hypothesis is based (Wilson et al., 1977) such as the assumption of a uniform rate of mutation may not be valid for extrapolations over very long periods of time.

References

Bernard, M. P., Myers, J. C., Chu, M.-L., Ramirez, F., Eikenberry, E. F., & Prockop, D. J. (1983) *Biochemistry 22*, 1139.

Bornstein, P., & Traub, W. (1979) *Proteins (3rd Ed.) 4*, 417.

Bornstein, P., & Sage, H. (1980) *Annu. Rev. Biochem. 49*, 957.

Carpousis, A., Christner, P., & Rosenbloom, J. (1977) *J. Biol. Chem. 252*, 8023.

Chu, M.-L., Myers, J. C., Bernard, M. P., Ding, G., & Ramirez, F. (1982) *Nucleic Acids Res. 10*, 5925.

Clifford, A. A. (1973) *Multivariate Error Analysis*, Halsted Press, New York.

Cloud, P. E. (1968) *Science (Washington, D.C.) 160*, 729.

Dickson, L. D., Ninomiya, Y., Bernard, M. P., Pesciotta, D. M., Parsons, J., Green, G., Eikenberry, E. F., de Crombrugghe, B., Vogeli, G., Pastan, I., Fietzek, P. P., & Olsen, B. R. (1981) *J. Biol. Chem. 256*, 8407.

Dixit, S. N., Seyer, J. M., & Kang, A. H. (1977a) *Eur. J. Biochem. 73*, 213.

Dixit, S. N., Seyer, J. M., & Kang, A. H. (1977b) *Eur. J. Biochem. 81*, 599.

Dixit, S. N., Mainardi, C. L., Seyer, J. M., & Kang, A. H. (1979) *Biochemistry 18*, 5417.

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slighton, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., & Proudfoot, N. J. (1980) *Cell (Cambridge, Mass.) 21*, 653.

Evans, R. D. (1955) *The Atomic Nucleus*, p 780, McGraw-Hill, New York.

Fietzek, P. P., & Kühn, K. (1973) *FEBS Lett. 36*, 289.

Fietzek, P. P., Breitkreutz, D., & Kühn, K. (1974a) *Biochim. Biophys. Acta 365*, 205.

Fietzek, P. P., Furthmayr, H., & Kühn, K. (1974b) *Eur. J. Biochem. 47*, 257.

Fuller, F., & Boedtker, H. (1981) *Biochemistry 20*, 996.

Hanukogu, I., Tanese, N., & Fuchs, E. (1983) *J. Mol. Biol. 163*, 673.

Highberger, J. H., Corbett, C., Kang, A. H., & Gross, J. (1975) *Biochemistry 14*, 2872.

Highberger, J. H., Corbett, C., Dixit, S. N., Yu, W., Seyer, J. M., Kang, A. H., & Gross, J. (1982) *Biochemistry 21*, 2048.

Hofmann, H., Fietzek, P. P., & Kühn, K. (1978) *J. Mol. Biol. 174*, 137.

Kang, A. H., & Gross, J. (1970) *Biochemistry 9*, 796.

Kimura, M., & Ohta, T. (1972) *J. Mol. Evol. 2*, 87.

Martin, S. L., Zimmer, E. A., Davidson, W. S., & Wilson, A. C. (1981) *Cell (Cambridge, Mass.) 25*, 737.

Mathews, M. B. (1980) in *Biology of Collagen* (Viidik, A., & Vuust, J., Eds.) p 193, Academic Press, New York.
Maxam, A., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499.
McKenna, M. C. (1975) in *Phylogeny of the Primates* (Luckett, W. P., & Szalay, F. S., Eds.) p 21, Plenum Press, New York.
Monson, J. M., Friedman, J., & McCarthy, B. J. (1982) *Mol. Cell. Biol. 2*, 1362.
Myers, J. C., Chu, M.-L., Faro, S. H., Clark, W. J., Prockop, D. J., & Ramirez, F. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 3516.
Olsen, B., & Dickson, L. (1981) in *The Chemistry and Biology of Mineralized Connective Tissues* (Veis, A., Ed.) p 143, Elsevier/North-Holland, New York.
Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., & Dodgson, J. (1980) *Cell (Cambridge, Mass.) 20*, 555.
Piez, K. A. (1976) in *Biochemistry of Collagen* (Ramachandran, G. N., & Reddi, A. H., Eds.) p 1, Plenum Press, New York.
Piez, K. A. (1980) in *Gene Families of Collagen and Other Proteins* (Prockop, D. J., & Champe, P. C., Eds.) p 142, Elsevier/North-Holland, New York.
Prockop, D. J., Kivirikko, K. I., Tuderman, L., & Guzman, N. A. (1979) *N. Engl. J. Med. 301*, 13, 77.
Showalter, A. M., Pesciotta, D. M., Eikenberry, E. F., Yamamoto, T., Pastan, I., de Crombrugghe, B., Fietzek, P. P., & Olsen, B. R. (1980) *FEBS Lett. 111*, 61.
Wilson, A. C., Carlson, S. S., & White, T. J. (1977) *Annu. Rev. Biochem. 46*, 573.
Wozney, J., Hanahan, D., Tate, V., Boedtker, H., & Doty, P. (1981) *Nature (London) 294*, 129.
Yamamoto, T., Sobel, M. E., Adams, S. L., Avvedimento, V. E., DiLauro, R., Pastan, I., de Crombrugghe, B., Showalter, A., Pesciotta, D. M., Fietzek, P. P., & Olsen, B. R. (1980) *J. Biol. Chem. 255*, 2612.

Enzymes of the β-Ketoadipate Pathway in *Pseudomonas putida*: Kinetic and Magnetic Resonance Studies of the *cis,cis*-Muconate Cycloisomerase Catalyzed Reaction†

Ka-Leung Ngai, L. Nicholas Ornston, and Roland G. Kallen*

ABSTRACT: Steady-state kinetic analysis of the divalent metal ion requiring *cis,cis*-muconate cycloisomerase catalyzed interconversion of *cis,cis*-muconate and (+)-muconolactone obeys Michaelis–Menten kinetics and the Haldane relationship from pH 6.2 to 8.3. The pH vs. k_{cat}/K_m profiles suggest free-enzyme apparent pK_a values of 6.2 and 7.4: the reciprocal behavior of the data with respect to the latter pK_a value is consistent with base–acid catalysis by the enzyme involving proton removal from the lactone and protonation of *cis,cis*-muconate, respectively. This catalysis by the enzyme of proton transfer is consistent with the stereospecific incorporation of solvent deuterium into the *pro*-5*R* position of (+)-muconolactone in the enzyme-catalyzed reaction: in reverse, the departure of the carboxylic oxygen atom and proton from the C(4) and C(5) carbon atoms follows a syn (cis) route [Avigad, G., & Englard, S. (1969) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 28*, 345, Abstr. 486]. The titration of enzyme freed of divalent metal ion with manganous ion, monitored by electron paramagnetic resonance spectroscopy and steady-state kinetic measurements, indicates a single binding site per subunit characterized by $K_{diss}^{E\cdot Mn} = [E][Mn^{2+}]/[E\cdot Mn^{2+}] = 4.5$ and 3.0 μM, respectively, the latter value analyzed via a rapid equilibrium mechanism. The paramagnetic effects of Mn^{2+} on the $1/T_1$ and $1/T_2$ values for the H-5*S* proton of (+)-muconolactone in the E·ML·Mn ternary complex provide an estimate of the correlation time, τ_c, at 5×10^{-9} s from the T_1/T_2 ratio, indicating that the condition of rapid exchange of (+)-muconolactone in solution with the ternary complex obtains. From the T_2 values, the rate constant for dissociation of (+)-muconolactone from the ternary complex is 3.3×10^4 s^{-1}, which is greater than the catalytic center activity for studies of the reaction in the direction *cis,cis*-muconate to (+)-muconolactone by a factor of about 10^3, consistent with the rapid equilibrium assumption. The distances from the enzyme-bound Mn^{2+} to the H-2, H-3, and H-5*S* protons calculated from the T_1 values for the individual protons are 5.8, 5.3, and 5.2 Å, respectively. The distance from the enzyme-bound Mn^{2+} to the H-5*S* proton of (+)-muconolactone in the ternary complex appears to allow either direct or outer sphere coordination of the metal ion to the C(6)-carboxyl group of (+)-muconolactone. However, EPR spectra of the E·Mn·ML complex at 35 GHz are identical with spectra of the E·Mn complex, making unlikely direct coordination and suggesting that, alternatively, the metal ion may merely serve a structural role.

cis,cis-Muconate cycloisomerase[1] (EC 5.5.1.1) is one of the enzymes involved in the catabolism of benzoic acid by means of the pyrocatechol branch of the β-ketoadipate pathway in *Pseudomonas putida*. This enzyme catalyzes the reversible

† From the Department of Biochemistry and Biophysics, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104 (K.-L.N. and R.G.K.), and the Department of Biology, Yale University, New Haven, Connecticut 06520 (L.N.O.). *Received January 14, 1983; revised manuscript received May 24, 1983.* Supported by Grant GM 13777 from the National Institutes of Health (R.G.K.), Grants GK 27697K (R.G.K.) and PCM 77-24834 (L.N.O.) from the National Science Foundation, Grant RR 00512 from the Middle Atlantic NMR Research Facility, and Research Career Development Award K4 CA 70487 from the National Cancer Institute (R.G.K.). A portion of this work has been published in preliminary form (Ngai & Kallen, 1977, 1979), and further details may be found in a Ph.D. Dissertation (Ngai, 1981).

[1] Abbreviations: CCM, *cis,cis*-muconate (2,4-hexadienedioic acid); E·CCM, ternary complex of enzyme, metal ion, and CCM; E·ML, ternary complex of enzyme, metal ion, and ML; $[E_t]$, total enzyme concentration on a subunit basis; DOD, deuterium oxide; DSS, 4,4-dimethyl-4-silapentane-1-sulfonic acid; EDTA, ethylenediaminetetraacetic acid; ML, (+)-muconolactone [γ-(carboxymethyl)-Δ^{α}-butenolide; 4-(*S*)-hydroxy-2-hexenedioic acid 1,4-lactone]; Hepes, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid; Mes, 4-morpholineethanesulfonic acid; Pipes, 1,4-piperazinediethanesulfonic acid; Tris, tris(hydroxymethyl)aminomethane; Hepps, 4-(2-hydroxyethyl)-1-piperazinepropanesulfonic acid; EPR, electron paramagnetic resonance; NMR, nuclear magnetic resonance. *cis,cis*-Muconate cycloisomerase has also been referred to as *cis,cis*-muconate lactonizing enzyme (MLE).

0006-2960/83/0422-5223$01.50/0