Fibrillar Collagen Genes

Structure and Expression in Normal and Diseased States[a]

FRANCESCO RAMIREZ,[b] SHARON BOAST,
MARINA D'ALESSIO, BRENDAN LEE,
JAMES PRINCE, MING-WAN SU,
HENRIK VISSING, AND
HIDEKATSU YOSHIOKA

Department of Microbiology and Immunology and
Department of Orthopedic Surgery
Morse Institute of Molecular Genetics
State University of New York
Health Science Center at Brooklyn
Brooklyn, New York 11203

INTRODUCTION

The collagens represent a multigene family whose members are segregated into evolutionarily distinct groups. The largest of these groups (group 1) includes the genes that specify the procollagen precursors of the fibril-forming molecules (types I, II, III, V, and XI). To date seven of the group 1 genes have been cloned, and their corresponding loci have been mapped. Despite some important differences, the overall organization of the fibrillar collagen genes is highly similar and remarkably conserved in Deuterostomia. This substantiates the hypothesis that the group 1 collagens arose from a common multi-exon progenitor prior to vertebrates' radiation. The characterization of several collagenopathies has elucidated some of the selective pressures that have rigidly maintained such a complex organization during evolution. These studies have also defined the metabolic consequences of collagen mutations and the unique contribution of each collagen type to the integrity of the body architecture. In contrast to their structural similarities, the differential expression of the group 1 genes correlates with the presence of distinct networks of interactions between cis-acting sequences, as well as between these elements and specific *trans*-acting nuclear factors. Relevant to this, the finding that fibrillar collagen genes are activated during

[a]This work was supported by grants from the National Institutes of Health (AR 38648, HD 22657, and HL 41104) and the March of Dimes Birth Defects Foundation (1-1042).

[b]Present address: Brookdale Center for Molecular Biology, Mt. Sinai Medical Center, 1 Gustave Levy Place, New York, New York 10029.

early animal embryogenesis supports the notion that this ubiquitous and ancient group of proteins plays an active role in critical morphogenetic programs of all metazoa, as well.

Our work on these three topics (gene structure and evolution; pathogenesis and physiology; regulation and developmental expression) will be briefly reviewed here, with particular emphasis on the human system.

GENE STRUCTURE AND EVOLUTION

A fibrillar procollagen molecule can be generalized as consisting of three major domains, a long triple-helical region made of several Gly-X-Y repeats and two short terminal projections.[1]

The structure of the C-terminal propeptide is fairly similar in all procollagen chains, and it is characterized by the presence of highly conserved elements, notably the C-protease cleavage site, the cysteinyl residues, and the N-linked carbohydrate attachment site. C-propeptides play a key metabolic function since intracellular folding of procollagens into the triple helix begins with the formation of inter- and intrachain disulfide bonds.[1] The spatial organization of the cysteinyl residues thought to be involved in these interactions is seemingly invariant among fibril-forming procollagens; to be specific, the first four residues (C1 to C4) are believed to participate in interchain linkages, the remaining (C5 to C8) in intrachain bonding.[1] The pattern of the first four cysteines, however, displays some exceptions in that pro-α2(I) and pro-α1(XI) lack C2, and pro-α2(V) lacks C3.[2-5] Interestingly, the C-propeptides of two echinoid fibrillar collagens exhibit the same structural features as their vertebrate counterparts, thus suggesting that intra- and intermolecular junctions, as well as N-linked glycosylation, are fundamental structural/functional elements of the molecular architecture in vertebrate and invertebrate species.[6-7] At the gene level, the C-propeptide coding sequences are similarly divided into four exons (numbers 49 to 52), the first of which contains both collagenous and C-propeptide sequences.[8]

With the exception of two fused exons in COL1A1, the triple-helical coding exons of the group 1 genes conform to an organizational pattern in which both their number and size are rigidly maintained.[2,4,5,8,9] The exons are mostly 54 bp in length or multiples of it (45 bp, 99 bp, 108 bp, 162 bp). This strongly suggests that the triple-helical coding domain evolved by processes of duplication, fusion, and deletion of an ancestral 54-bp unit that codes for six Gly-X-Y triplets, or 1½ turns of the helix.[10] Relevant to this, it has recently been found that a bacterial hydrolase is stabilized into a homotrimeric conformation by six Gly-X-Y repeats, as well.[11] Irrespective of individual sizes, each of the collagenous coding exons begins with a Gly codon and ends with a Y codon, thus resulting in an in-frame "cassette" type of organization. Such a motif has recently been identified in sea urchin collagen genes but not in those of *Drosophila melanogaster* and *Caenorhabditis elegans*.[6,7,12-15] Hence, the different organization of the invertebrate collagen genes seems to reflect more fundamental differences in the evolution of the Protostomia and Deuterostomia genomes.

In contrast to the other two domains, the N-propeptides differ greatly in size and composition among group 1 chains and appear to adhere to one of two basic architectures.[15] The first, characteristic of pro-α1(I), pro-α1(III), and pro-α2(V), displays the longitudinal arrangement of a cysteine-rich globular domain, a collagenous sequence, and a short, nonhelical segment.[9,17,18] The second, characteristic of pro-α2(I)

and pro-α1(II), lacks the entire cysteine-rich globular domain, and, as a consequence, the signal peptide is directly connected to the collagenous sequence.[2,19] More recently, a third architecture has been recognized in pro-α1(XI), whose N-propeptide is significantly longer than the others; it contains short, dispersed clusters of collagenous sequences and lacks the cysteine-rich globular domain (H. Yoshioka *et al.*, manuscript in preparation). It should be noted that interrupted collagenous sequences are also observed in pro-α2(V) and pro-α1(II).[17,19] Interestingly, a phylogenetic comparison of the latter chain in man, chicken, and frog has revealed that the loss of the cysteine-rich globular region is a relatively recent event, for this domain is present in the amphibian, but not in the avian and human chains (M. W. Su *et al.*, manuscript in preparation). At the genomic level the N-propeptide coding sequences of COL1A1, COL1A2, COL2A1, and COL3A1 are organized in distinct manners with consequent divergency in both the number and the size of the corresponding exons.[2,9,18,19] The numerical heterogeneity of the N-terminal coding units seems, at the present time, to correlate more with the heterotrimeric assembly of their products than with the particular architecture of individual N-propeptides.[19]

From the data hitherto gathered, it is therefore clear that the evolutionary history of the group 1 collagen genes is ancient, in that it predates vertebrates' radiation. In contrast to the triple-helical domain, whose 54-bp exon seems to have been selected as an optimal unit for chain interaction, the N- and C-propeptides seem to have followed distinct and more relaxed evolutionary pathways. The detail of such a scenario is likely to be elucidated when the organization of additional members of the group 1 family is characterized in different metazoa.

PATHOGENESIS AND PHYSIOLOGY

Because of the metabolic, structural, and genomic similarities of the group 1 collagens, the well-characterized type I collagenopathies serve as paradigms for the analysis of other connective tissue disorders. To date two clinically distinct conditions are associated with mutations in the subunits of the type I collagen, namely osteogenesis imperfecta (OI) and Ehlers-Danlos syndrome Type VII (EDS VII).[20] In broad terms the former, characterized by bone fragility, results from mutations that affect the intracellular assembly of the trimers; whereas the latter, characterized by loose joints, is caused by defects that interfere with extracellular conversion of procollagen to collagen.

OI mutations, which include deletions, mis-splicing, and single amino acid substitutions, are all believed to decrease the rate of helical assembly and thus to expose greater regions of unassembled chains to overmodification.[21] Hence, the relative location of the defect within the triple-helical domain is thought to determine the overall rate of modification and consequently the degree of phenotypical severity.[21] OI mutations have also provided some relevant information on the evolution of the genes. First, the essential role of the highly conserved cysteinyl residues of the C-propeptide has been demonstrated by the finding that loss of C8 in pro-α2(I) collagen leads to the exclusion of this chain from trimer assembly.[22] Second, the absolute preservation of the organization of the triple-helical coding domain has been inferred from deletion mutants in which shortened but in-frame procollagen chains lead to the formation of out-of-phase parental trimers that are incompatible with survival.[23]

Some exceptions to the aforementioned rules are now being discovered in type II and type III collagenopathies. In two EDS VII patients, large heterozygous deletions

in the triple-helical domain of COL3A1 call into question the validity of the so-called location rule, since they exhibit an inverse relationship between clinical severity and C-terminal proximity of the mutations.[24–27] Similarly, the finding that a single-exon deletion and a Gly-to-Ser substitution at nearly the same location in α1(II) lead to severe (spondyloepiphyseal displasia) and lethal phenotypes (type II achondrogenesis-hypochondrogenesis), respectively, implies that the nature of the mutation may also play an important role.[28–30]

EDS VII has provided novel information on the mechanisms of splice-site selections in multi-exon genes. The five heterozygous cases hitherto studied display in fact the same splicing abnormality, exon-skipping, which is, however, caused by three distinct mutations. The first, which involves the obligatory GT dinucleotide of the 5′ splice site of the sixth intron of COL1A2, leads to precise outsplicing of exon-6 sequences with the consequent elimination of the N-proteinase cleavage site and the N-telopeptide cross-linking site.[31] A similar event is seen in the COL1A2 allele of another EDS VII patient,[32] in which all of the sequences believed to regulate splicing are surprisingly normal. Conceivably, the mutation resides within the large intervening sequence that flanks exon 6, thus suggesting that splicing of such a complex transcript might be dependent on as-yet-uncharacterized elements. The remaining three EDS VII variants bear the same mutation in either COL1A1 or COL1A2, notably a G-to-A substitution in the last codon of exon 6.[33–34] Such a finding reiterates the importance of the whole exon/intron junction sequence in conferring specificity to the splicing reaction. The exon change decreases but does not abolish the relative strength of the splicing signal, in that corrected spliced transcripts are demonstrable in both the COL1A2 (50%) and COL1A1 (20%) mutated alleles.[33–34] In line with this and in contrast to the intron mutation, the mis-splicing caused by the exon mutation can be effectively suppressed *in cellula* by lowering the temperature at which the fibroblasts are grown.[34] It can also be argued, by analogy to *in vitro* splicing experiments,[39] that the differential effectiveness of the exon 5/intron 5 junction to compete in *cis* for splice-site recognition in COL1A2 vs. COL1A1 may reflect the difference in relative distance of the splicing signals in the two genes.[2,9] Accordingly, we suggest that splicing mutations in collagen genes can have variable expression depending on which of the sequences are affected, on the relative strength and distance of the *cis*-competing signal, and plausibly on exogenous factors such as temperature, which may modulate mis-splicing in different tissues and/or at different developmental stages.

REGULATION AND DEVELOPMENTAL EXPRESSION

The sequences around the transcription start site of human type I collagen genes have been functionally analyzed in search of *cis*-acting elements that regulate optimal and tissue-specific expression.

We have found that the COL1A2 promoter (−3800 +54), linked to a reporter gene—the bacterial chloroamphenicol acetyltransferase (CAT)—and transfected into human fibroblasts and lymphoblasts, displays tissue-specific expression. A series of CAT constructs containing 5′ COL1A2 deletions identified the region necessary for proper transcription from position −772 to +54 (S. Boast *et al.*, manuscript in preparation). In the same experiments we also demonstrated that the type I and type II collagen genes possess distinct regulatory elements, in that CAT activity of the minimal COL1A2 promoter is effectively abolished by competition with the promoter sequences of COL1A1 but not COL2A1. When COL1A2-driven CAT activity was competed by internal fragments of the COL1A2 promoter, inhibition was observed

with all competing molecules in fibroblasts. In parallel experiments utilizing immortalized chondrocytes as recipient cells, all but one of the COL1A2 fragments (−104 +54) displayed efficient competition of the basal CAT activity. This finding suggests that chondrocytes lack fibroblasts' specific factor(s) binding to the sequences between −104 +54 of COL1A2. In a SV40-transformed human fibroblast cell line, which does not produce pro-α2(I) collagen,[35] the COL1A2/CAT construct was found to mimic the downregulation of the endogenous gene when transfected at low concentration.[37] At higher concentrations of the chimeric construct CAT activity was fully restored, indicating the presence of (a) titratable transacting factor(s) present only in the nuclear lysate of the transformed cells.[37] Competition experiments and band-shift assays revealed that binding of the factor(s) occurs in the region of COL1A2 comprised between −360 and +54.[37] Hence, most of the sequences regulating its expression appear to reside within this short segment of the COL1A2 promoter.

Similar experiments have demonstrated the presence of a downstream *cis*-acting enhancer element in COL1A1. Briefly, transcriptional activity of the COL1A1 promoter was originally assessed on an α globin gene after injection of the construct into *Xenopus laevis* oocytes or transfection into NIH 3T3 cells. When the activity of this reference construct was compared to that of a similar one that contains in addition 782 bp of the first intron of COL1A1, a significant enhancement in globin production was observed.[38] Subsequently, however, others have shown that in chicken embryo fibroblasts the COL1A1 intervening sequence fails to enhance transcription.[39–41] To reconcile this discrepancy, the activity of the 782-bp element positioned in both orientations upstream from a COL1A1 promoter/CAT construct has recently been tested in a variety of cells, including human fibroblasts, human lymphoblasts, and chicken embryo fibroblasts (S. Boast *et al.*, manuscript in preparation). The results showed that the 782-bp element enhances—in a preferential but not exclusive orientation—its own promoter without tissue specificity but rather with species specificity, for it does not display activity in avian cells.

To complement and extend our study on collagen regulation the pattern of collagen onset during the development of sea urchin and *Xenopus laevis* embryos has recently been ascertained. In the former organism two fibrillar collagen genes were found to be expressed in the mesenchyme cell lineage during early stages of embryogenesis.[6,7] On the basis of these data, we concluded that the two genes specify molecules likely to be involved in both gastrulation and skeletogenesis.[42] Likewise, in the frog's embryo type II collagen transcripts were detected at the time of mesodermal induction,[42] thus suggesting that this fibrillar collagen is probably involved in the spatial organization of the developing embryo prior to organogenesis (M. W. Su *et al.*, manuscript in preparation). It will be of interest to further elucidate the function and regulation of collagen in these two organisms in which activation of cell lineage-specific genes results from either determinative or inductive responses.

In conclusion, the analysis of the biology of the fibrillar collagen genes in vertebrate and invertebrate species, as well as in pathological and physiological conditions, has demonstrated the unique versatility of this system in addressing a variety of questions that greatly impact on our understanding of eukaryotic gene expression and evolution.

ACKNOWLEDGMENTS

The authors thank Ms. R. Lingeza for preparing the manuscript.

REFERENCES

1. MILLER, E. J. 1985. Ann. N.Y. Acad. Sci. **460:** 1-13.
2. DE WET, W., M. BERNARD, V. BENSON-CHANDA, M. L. CHU, L. DICKSON, D. WEIL & F. RAMIREZ. 1987. J. Biol. Chem **262:** 16032-16036.
3. DION, A. S. & J. C. MEYERS. 1987. J. Mol. Biol. **193:** 127-143.
4. WEIL, D., M. BERNARD, S. GARGANO & F. RAMIREZ. 1987. Nucleic Acids Res. **15:** 181-198.
5. BERNARD, M., H. YOSHIOKA, E. RODRIQUEZ, M. VAN DER REST, J. KIMURA, Y. NINOMIYA, B. R. OLSEN & F. RAMIREZ. 1988. J. Biol. Chem. **263:** 17159-17166.
6. D'ALESSIO, M., F. RAMIREZ, H. SUZUKI, M. SOLURSH & R. GAMBINO. 1989. Proc. Natl. Acad. Sci. USA. In press.
7. D'ALESSIO, M., F. RAMIREZ, H. SUZUKI, M. SOLURSH & R. GAMBINO. 1990. Ann. N.Y. Acad. Sci. This volume.
8. RAMIREZ, F., M. BERNARD, M. L. CHU, L. DICKSON, F. SANGIORGI, D. WEIL, W. DE WET, C. JUNIEN & M. E. SOBEL. 1985. Ann. N.Y. Acad. Sci. **460:** 117-124.
9. CHU, M. L., W. DE WET, M. BERNARD, J. F. DING, M. MORABITO, J. MYERS, C. WILLIAMS & F. RAMIREZ. 1984. Nature **310:** 337-340.
10. VOGELI, G. H., H. OHKUBO, V. E. AVVEDIMENTO, M. SULLIVAN, Y. YAMADA, J. M. MUDRYJ, I. PASTAN & B. DE CROMBRUGGHE. 1980. Cold Spring Harbor Symp. Quant. Biol. **45:** 777-783.
11. CHARALAMBANS, B. M., J. N. KEEN & M. J. MCPHERSON. 1988. EMBO J. **7:** 2903-2909.
12. VENKATESAN, M., F. DE PABLO, G. VOGELI & R. T. SIMPSON. 1986. Proc. Natl. Acad. Sci. USA **83:** 3351-3355.
13. SAITTA, B., G. BUTTICE & R. GAMBINO. 1989. Biochem. Biophys. Res. Commun. **158:** 633-639.
14. MONSON, J. M., J. NATZLE, J. FRIEDMAN & W. MCCARTHY. 1982. Proc. Natl. Acad. Sci. USA **79:** 1761-1765.
15. KRAMER, J. M., G. N. COX & D. HIRSH. 1982. Cell **30:** 599-606.
16. TIMPLE, R. & R. W. GLANVILLE. 1981. Clin. Orthop. Relat. Res. **158:** 224-242.
17. WOODBURY, D., V. BENSON-CHANDA & F. RAMIREZ. 1989. J. Biol. Chem. **264:** 2735-2738.
18. BENSON-CHANDA, V., M. W. SU, D. WEIL, M. L. CHU & F. RAMIREZ. 1989. Gene. **78:** 255-265.
19. SU, M. W., V. BENSON-CHANDA, H. VISSING & F. RAMIREZ. 1989. Genomics **4:** 438-441.
20. TSIPOURAS, P. & F. RAMIREZ. 1987. J. Med. Genet. **24:** 2-8.
21. BONADIO, J. & P. BYERS. 1985. Nature **316:** 363-366.
22. PIHLAJANIEMI, T. L., L. A. DICKSON, F. M. POPE, V. R. KORHONER, A. NICHOLLS, D. J. PROCKOP & J. MYERS. 1984. J. Biol. Chem **259:** 12941-12944.
23. CHU, M. L., V. GARGIULO, C. J. WILLIAMS & F. RAMIREZ. 1985. J. Biol. Chem. **260:** 691-694.
24. SUPERTI-FURGA, A., E. GUGLER, R. GITZELMANN & B. STEINMANN. 1988. J. Biol. Chem **263:** 6226-6232.
25. SUPERTI-FURGA, A., B. STEINMANN, F. RAMIREZ & P. H. BYERS. 1989. Hum. Genet. **82:** 104-108.
26. SUPERTI-FURGA, A., H. VISSING, B. LEE, F. RAMIREZ, P. BYERS & B. STEINMANN. 1989. Submitted for publication.
27. SUPERTI-FURGA, A. & B. STEINMANN. 1988. Biochem. Biophys. Res. Commun. **150:** 140-147.
28. LEE, B., H. VISSING, F. RAMIREZ, D. ROGERS & D. RIMOIN. 1989. Science **244:** 978-980.
29. GODFREY, M. & D. HOLLISTER. 1988. Am. J. Hu. Genet. **43:** 904-913.
30. VISSING, H., M. D'ALESSIO, B. LEE, F. RAMIREZ, M. GODFREY & D. HOLLISTER. 1989. J. Biol. Chem. In Press.
31. WEIL, D., M. BERNARD, N. COMBATES, M. K. WIRTZ, D. HOLLISTER, B. STEINMANN & F. RAMIREZ. 1988. J. Biol. Chem. **263:** 8561-8564.
32. EYRE, D. R., F. D. SHAPIRO, J. F. ALDRIDGE. 1985. J. Biol. Chem. **260:** 11322-11329.
33. WEIL, D., M. D'ALESSIO, F. RAMIREZ, W. DE WET, W. G. COLE & J. F. BATEMAN. 1989. EMBO J. **8:** 1705-1710.

34. WEIL, D., M. D'ALESSIO, F. RAMIREZ, B. STEINMANN, M. K. WIRTZ, R. W. GLANVILLE & D. W. HOLLISTER. 1989. J. Biol. Chem. In Press.
35. REED, R. & T. MANIATIS. 1986. Cell **46:** 681-690.
36. PARKER, M. I., A. A. SMITH & W. GEVERS. 1989. J. Biol. Chem. **264:** 7147-7152.
37. PARKER, M. I., A. A. SMITH, S. BOAST, M. W. SU & F. RAMIREZ. 1990. Ann. N.Y. Acad. Sci. This volume.
38. ROSSOUW, C. M. S., S. J. DU PLOOY, M. BERNARD, F. RAMIREZ & W. DE WET. 1987. J. Biol. Chem. **262:** 15151-15157.
39. BORNSTEIN, P., J. MCKAY, J. K. MORISHIMA, S. DEVARAYALU & R. GELINAS. 1987. Proc. Natl. Acad. Sci. USA **84:** 8864-8868.
40. BORNSTEIN, P. & J. MCKAY. 1988. J. Biol. Chem. **263:** 1603-1606.
41. BORNSTEIN, P., J. MCKAY, D. J. LISKA, S. APONE & S. DEVARAYALU. 1988. Mol. Cell. Biol. **8:** 4851-4857.
42. DAVIDSON, E. H. 1986. Gene Activity in Early Development. 3rd edit. Academic Press. New York, N.Y.