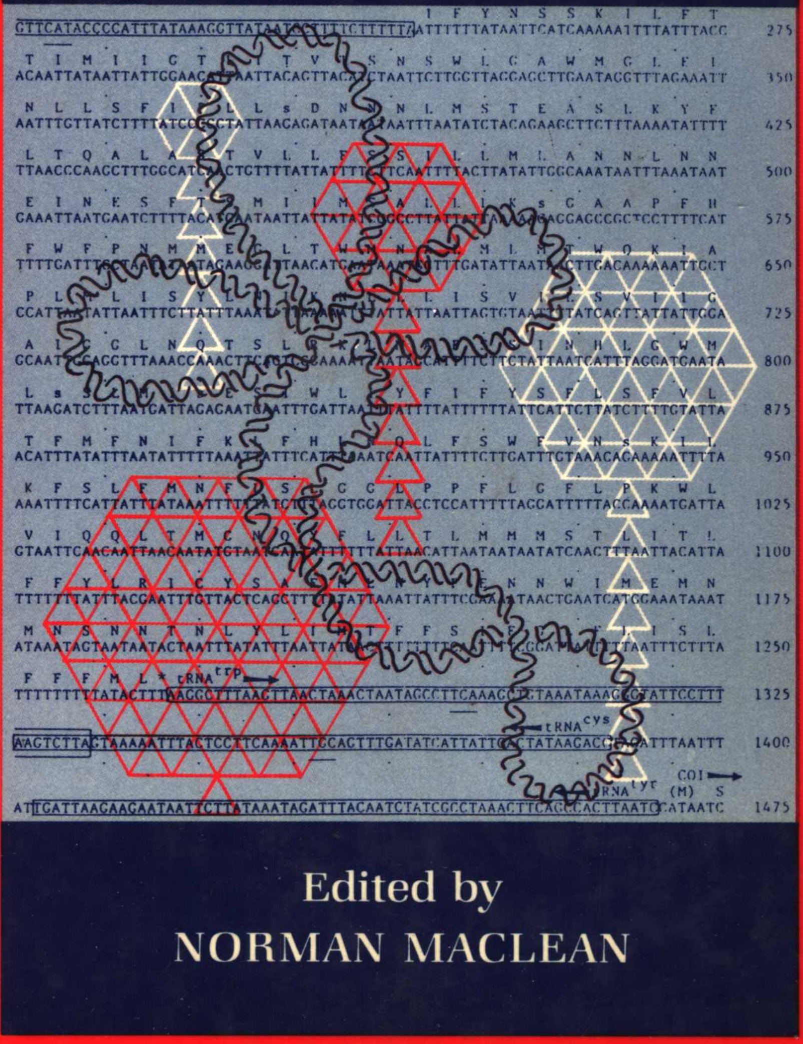
Oxford Surveys on Eukaryotic Genes

VOLUME 4

1987



Edited by

NORMAN MACLEAN

OXFORD SURVEYS ON EUKARYOTIC GENES

EDITED BY

NORMAN MACLEAN

VOLUME 4

1987

OXFORD UNIVERSITY PRESS

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Kuala Lumpur Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland
and associated companies in
Beirut Berlin Ibadan Nicosia

British Library Cataloguing in Publication Data
Oxford surveys on eukaryotic genes.
Vol. 4 (1987) –
1. Genetics 2. Eukaryotic cells
574.87'322 QH430

ISBN 0-19-854231-3

Filmset by Eta Services (Typesetters) Ltd, Beccles, Suffolk
Printed in Great Britain at the University Printing House, Oxford

Contributors

Ursula Bond: Department of Molecular Biophysics and Biochemistry, Yale University, Box 333, New Haven, Connecticut 06510, USA

Anamaris M. Colberg-Poley: Central Research and Development Department, E. I. du Pont de Nemours Co., Inc., Bldg. 328, Room B-22, Wilmington, Delaware 19898, USA

Peter Gruss: Department of Molecular Cell Biology, Max-Planck-Institut für biophysikalische Chemie, Am Fassberg, D-3400 Göttingen-Nikolausberg, FRG

Katherine Harding: Department of Biological Sciences, Fairchild Center, Columbia University, New York 10027, USA

Michael Levine: Department of Biological Sciences, Fairchild Center, Columbia University, New York 10027, USA

David W. Melton: Department of Molecular Biology, Edinburgh University, Mayfield Road, Edinburgh EH9 3JR, UK

Milton J. Schlesinger: Department of Microbiology and Immunology, Box 8093, Washington University Medical School, 660 South Euclid Avenue, St Louis, Missouri 63110, USA

James Scott: Division of Molecular Medicine, MRC Clinical Research Centre, Watford Road, Harrow HA1 3UJ, UK

Bryan Sykes: Nuffield Department of Pathology, John Radcliffe Hospital, Level 4, Headington, Oxon OX3 9DU, UK

Stephan D. Voss: Department of Human Oncology, Clinical Sciences Center, Room K4/449, University of Wisconsin Medical School, 600 Highland Avenue, Madison, Wisconsin 53792, USA

Malcolm Whiteway: Genetic Engineering Section, Biotechnology Research Institute, National Research Council of Canada, 6100, Avenue Royalmount, Montreal, Quebec H4P 2R2, Canada

Contents

1 Collagen gene structure

BRYAN SYKES

What is collagen?

An eminent developmental biologist recently said of collagen chemists and other analysts of the extracellular matrix, that they 'know everything and explain nothing' (Wolpert 1986). This light-hearted jibe nicely describes the predominant philosophy of the last two decades. The frantic and acquisitive search for factual information especially during the last ten years has created an ever enlarging and increasingly confusing family of collagens and collagen genes. This chapter, as well as fulfilling its main function of giving an up-to-date summary of this factual knowledge will introduce some order and make an attempt at some explanations; not, though, to the extent of those developmental biologists who might equally be made to answer the reciprocal charge that they explain everything and know nothing.

While introducing the subject of collagen it is normal practice to use phrases such as 'collagens comprise a family of closely related yet genetically distinct proteins that provide mechanical support in tissues'. This is a very misleading statement when applied to what is now regarded as the collagen family. It would be excessively mischievous to seriously question that some members of the collagen family, the fibrillar collagens, have stress-resisting mechanical support as a major function but there is absolutely no direct evidence that the other members of the family have anything to do with what is normally regarded as the collagens' traditional role. Rather, this role has been inferred by a process of muddled thought along the lines: Fibrillar collagens are stress-resisting→Fibrillar collagens contain triple-helix→Other collagens contain triple-helix→Therefore, other collagens are stress-resisting. This is just as invalid a syllogism as the following plainly absurd example: Cats have four legs→Cats hunt mice→Cows have four legs→Therefore cows hunt mice.

A more accurate definition of a collagen is an administrative one. Simply stated, collagens are discovered in collagen laboratories. Other molecules, which share the same features but are discovered elsewhere are not collagens. The structural feature linking all these molecules is the triple helix. It is this domain which is shared by all the collagens and by other proteins—C1q, acetylcholinesterase and pulmonary surfactant apoprotein—and it is this domain which forms the common thread to the chapter.

Since the concept of the collagen laboratory forms an important part of the administrative definition, this perhaps needs some expansion. Work on collagen really began when the histological entity of 'white fibrous tissue' was found to be composed more or less entirely of a single protein. Because tissues high in white fibrous tissue, especially bones, had always been the major component of gelatin and traditional glues they became known, as early as 1859, as collagenous [Greek *κoλλα* = *glue* + -gène = *-gen* (taken in the sense of 'producing')] and the protein, naturally enough, became collagen. Because it was an abundant and important natural resource a lot of the best early work on collagen was carried on in laboratories operating within industrial or quasi-industrial organizations. Since working with collagen required rather specialized techniques and skills, the 'collagen laboratory' was born. These few specialized laboratories between them generated almost all of what we now know about the protein.

Finding new collagens

In 1969, it was shown that collagen from cartilage was different from collagen from tendon (the traditional source). This important discovery sparked off an intensive search for other new collagens in these laboratories. The triple helix is very resistant to exopeptidases like pepsin and pronase so a major route for discovery of new collagens has been proteolytic digestion and examination of the resistant remnants. This process leaves behind not complete molecules but helical domains. The molecular structures from which these remnant domains originated are then painstakingly reconstructed—still an active process in some of the more recently discovered collagens.

In addition to protease resistance the other feature which distinguishes the helical domain is its amino acid sequence. In order to pack into the centre of the triple helix, every third residue is glycine, the only amino acid lacking a side-chain. There are other features of the sequence such as a high proline and aspartic acid content and the presence of hydroxyproline but it is the repetition of glycine at every third position that makes the sequence instantly recognizable. This has identified helical domains during the structural characterization of C1q, a subcomponent of the first component of the classical complement pathway, of the asymmetric form of acetylcholinesterase and of a family of mannose-binding proteins. Recently, the glycine repeats have been recognized in the translated sequence of a pulmonary surfactant apoprotein from genomic nucleotide sequencing. Though this demonstrates the potential of the polypeptide to form a helix, proof that it does so must await collagenase digestion and other characterizations.

The nucleotide homologies imposed by the sequence requirements for helix formation have opened up a new method of searching for these domains. Cross-hybridization between helix coding sequences is considerable and, at low stringency, has allowed 'collagen' clones to be picked from genomic libraries of *Drosophila*, the nematode *Caenorhabditis elegans* and the sea urchin *Strongylocentrotus purpuratus*, using vertebrate collagen sequences as hybridization probes. Sequencing confirmed the glycine repeats but, as previously, this is not formal proof of a helical domain, only an indi-

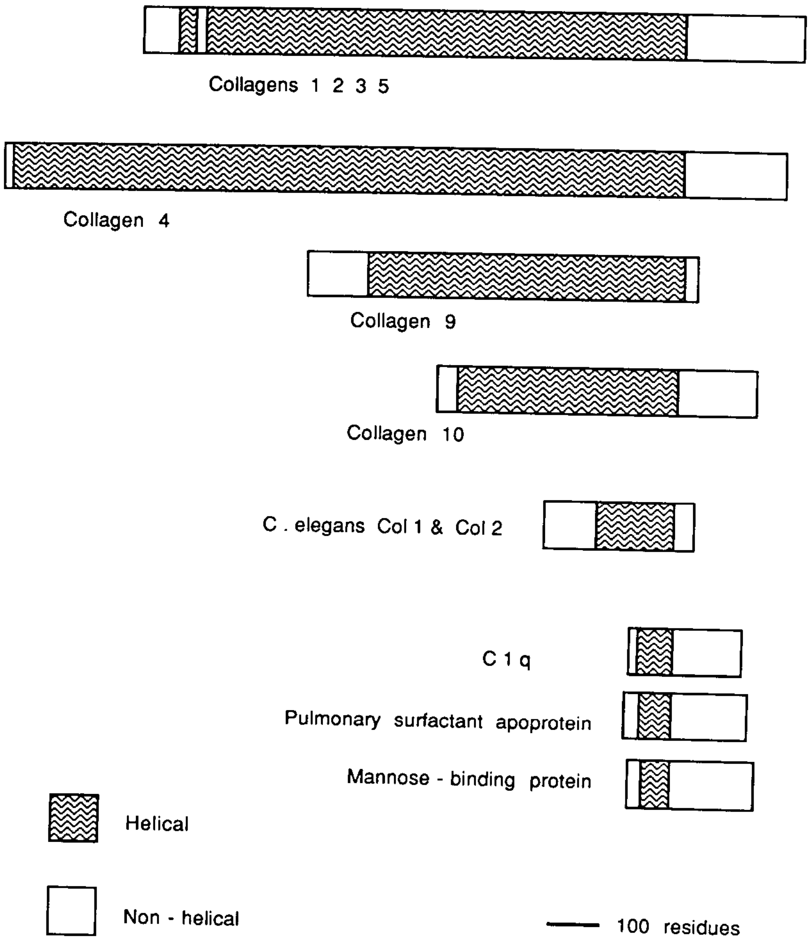


Fig. 1.1 Overall organization of the proteins considered.

cation of potential. If new members are to be added to the collagen family in the future, this is likely to be the major avenue of recruitment.

The net result of all this effort is a list of 15 vertebrate and four invertebrate proteins with evidence of (Gly–XY)$_n$ helical domains (Fig. 1.1).

The fibrillar collagens

Collagen 1 is by far the most abundant single protein in vertebrates. It is found in every tissue, except cartilage, and occurs as fibrils in the extracellular matrix. Negatively stained, these fibrils show a regular pattern of light and dark bands repeating every 67 nm. This is a consequence of the overlapping arrangement of the collagen 1 molecules within fibrils. Adjacent molecules are connected by covalent cross-links via short non-helical peptides (telopeptides) at either end. Extensive details of collagen synthesis, distribution, and molecular and fibrillar construction, are available in excellent reviews elsewhere (Bornstein and Sage 1980; Fessler *et al.* 1985). The implication of a mechanical function by all these features is obvious and is supported by direct evidence. This comes from three sources. Firstly, when the covalent crosslinks between collagen 1 molecules are either prevented from forming by lathyrogenic drugs or are disrupted by weak acid treatment, tissues lose resistance to mechanical stress and disintegrate. Secondly, when, in an experimental mouse strain, insertion of the murine Moloney retrovirus into the gene encoding the α1-chain of collagen 1 causes a complete transcriptional block, the result is a recessive embryonic lethal phenotype owing to rupture of major blood vessels (Lohler *et al.* 1984). Thirdly, biochemical and genetic linkage analysis of the inherited human disorder, osteogenesis imperfecta, shows that this phenotype in which tissues, especially bone, are unusually fragile, is caused by mutations in the genes encoding collagen 1 (Sykes *et al.* 1986). Interestingly, the mechanical stress to be resisted is not only externally applied forces but also the swelling pressure owing to the ability of matrix components to take up water. This can lead to a hundred-fold volume increase in tendons in which cross-links have been broken.

The similarity in molecular construction between collagen 1 and collagens 2, 3 and 5 and their ability to form fibrils strongly suggest they too have an engineering function*. Certainly, it is left to collagen 2 to resist the enormous swelling pressures generated by the highly-charged hydrophilic proteoglycans within cartilage. Whether collagens 3 and 5 form discrete homogeneous fibrils or participate in mixed fibril construction with collagen 1, is not clear, though there is evidence to support the latter suggestion. The fibrillar collagens and their genes are by far the best understood and their influence on thinking in the field cannot be overemphasized.

The mature collagen 1 molecule, as it occurs in crosslinked fibrils, is only a remnant of a longer precursor—procollagen—which loses substantial propeptides from both amino- and carboxy-termini after secretion into the matrix (Fig. 1.2). In common with other secreted proteins, procollagen is cleaved from an N-terminal hydrophobic signal peptide. Thus, the translation product of a fibrillar collagen gene encodes six distinct regions: (1) signal peptide, (2) N-terminal propeptide, (3) N-terminal telopeptide, (4) helical domain, (5) C-terminal telopeptide, (6) C-terminal propeptide. Of these, only domains 3, 4 and 5 are retained in the mature molecule.

As well as being the most abundant and best known of the fibrillar colla-

*The conventional notation of different collagens gives a Roman numeral to each 'type'. As the list of new collagen grows, this notation becomes clumsy, therefore use is made of Arabic numerals.

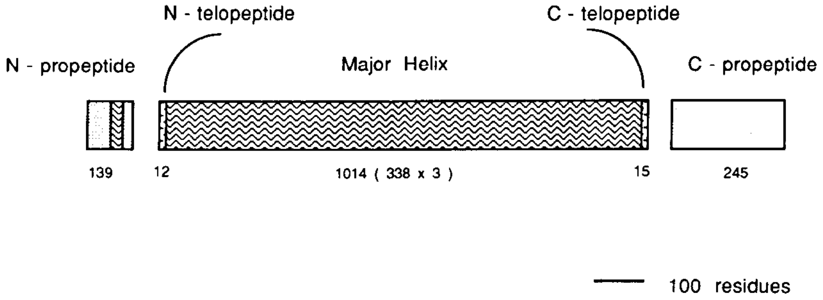


Fig. 1.2 Domains in procollagen I.

gens, collagen 1 is also interesting because it is composed of two different polypeptides, called α1 and α2 for historical chromatographic reasons. Each molecule contains two α1 chains and one α2 chain. Collagen 2 and 3 molecules on the other hand are each composed of identical chains. Collagen 5 appears to have three distinct chains which seem able to form trimeric molecules in different combinations. There is extensive sequence available on all fibrillar collagen genes. Because they are the most completely sequenced and illustrate all the major features and difficulties, the two genes encoding collagen 1 will be considered in detail, with reference being made to the other fibrillar genes when appropriate. The conventional locus nomenclature is the most convenient way of referring to genes encoding the different chains: COL1A1 encodes the collagen 1 (COL1) α1 chain (A1); COL1A2 encodes the collagen 1 α2 chain. Following this pattern COL2A1, COL3A1, COL5A1, COL5A2, COL5A3 encode the α1-chains for collagens 2, 3 and 5. The chick locus COL1A2 was the first to be completely sequenced and figures used here refer to this gene unless otherwise stated.

THE MAJOR HELICAL DOMAIN

The arrangement of sequences encoding the 1014 residue major helical domain is by far the most striking feature of the fibrillar collagen genes (Fig. 1.3). This domain is a perfect contiguous repeat of 338 (Gly–X–Y) triplets. The coding sequence for the domain is distributed between 44 exons. Apart from the two exons encoding triplet sequence at either end of the major helix, of which more will be described later, the 42 exons devoted entirely to encoding helical sequence have special features. Firstly, they all encode discrete numbers of triplets beginning at the 5′-end with the first G of a glycine codon and ending with the third base of a Y residue codon from the Gly–X–Y triplet. Secondly, there are only five different, yet related exon lengths. The most frequent (23/42) is 54 bp encoding 6 triplets; 108 bp (i.e. 54 × 2) exons occur eight times; exons of 45 bp (i.e. 54 − 9) and 99 bp (i.e. 108 − 9) each occur five times and there is one exon of 162 bp (54 × 3). It was

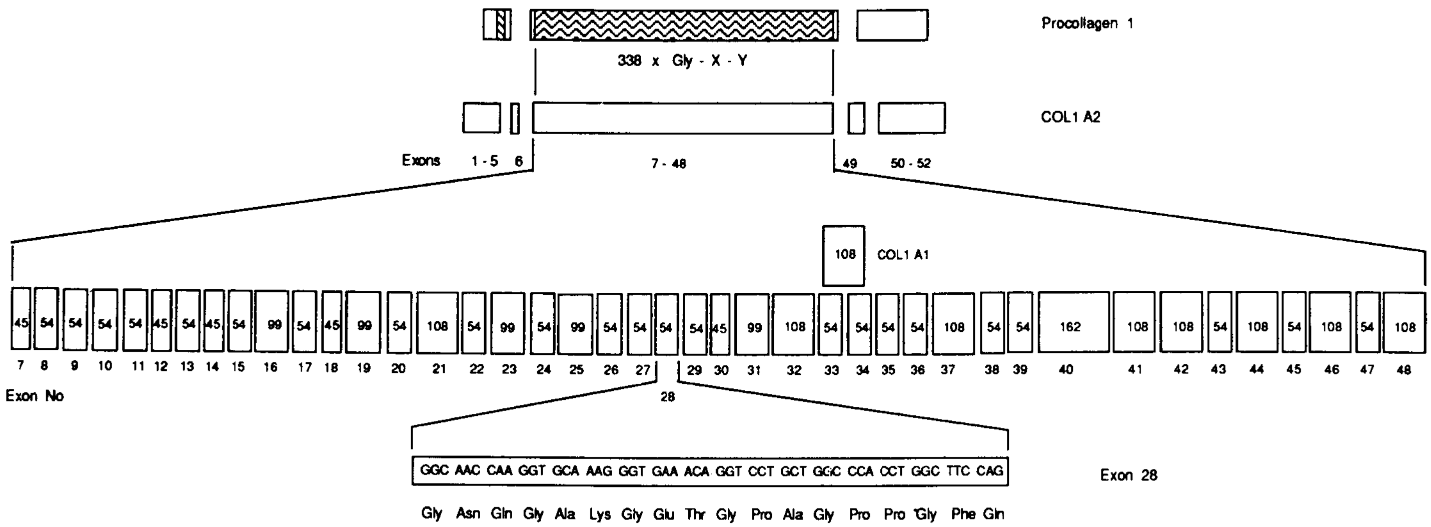


Fig. 1.3 Exon organization of the major helical domain of both COL1A1 and COL1A2. Numbering is from the 5′ end, so exon numbers will differ from most previous publications.

immediately obvious that derivation of the entire domain from a 54 bp ancestor by a mixture of a few simple manoeuvres was a possibility. An initial duplication of a 54 bp ancestor exon, itself probably the product of a six-fold reiteration of the basic GGNCCNCCN unit encoding a single Gly–Pro–Pro triplet, followed by further recombination between misaligned exons to give the 45, 99, 108 and 162 bp alternatives occurring at the same time as duplications achieved by recombination between introns. Then, suddenly, everything stopped.

This became clear as the exon arrangements of further fibrillar collagen genes were charaterized in more detail. It emerged that the order of exon sizes along the helical domain was precisely conserved not only between the same generic collagen from different species, which is not so unexpected, but also between genes encoding the different fibrillar collagens. There are a few minor exceptions. The two 54 bp exons 33 and 34 in COL1A2 are replaced by a single 108 bp exon in COL1A1. COL3A1 contains a further 18 bp of triplet sequence in exon 49 though this is a junction exon. Otherwise the exon order is precisely conserved.

The evidence for the evolution of all fibrillar collagen genes from a common ancestor is absolutely compelling. From rates of amino acid and nucleotide sequence substitution allowable within the constraints of absolute requirement for triplet conservation, the ancestors of the modern genes diverged between 500 and 1000 million years ago.

AN EVOLUTIONARY PARADOX

Most authorities have considered the evólution of the helical domain in fibrillar collagen genes as a straightforward process involving orderly duplications etc., of the 54 bp ancestral exon to the 'optimum' length followed by duplication then divergence of the modern lineages. This is quite impossible. The paradox is this: If the helical domain grew from a 54 bp ancestor by a complex series of recombination events there must be powerful forces at work to have prevented a continuation of the process since the lineages diverged. How, then, could these constraints, whatever they may be, have tolerated the process of helix expansion prior to divergence but not after it? It shall be argued that the constraints since divergence are functional and a hypothetical solution to the paradox will be suggested.

Whereas exon lengths and distribution are strictly conserved between helical domains of fibrillar collagen genes, intron lengths are not. The available information of overall helix domain length show a considerable variation between different genes ranging from 30 kb (chick COL1A2) to 11 kb (chick COL2A1). Neither is there conservation of intron length in the same gene between species. The helical domain of human COL2A1 is, at about 25 kb, over two-and-a-half times longer than the same domain in the chick. Since exon distribution is conserved it follows that these differences are entirely accounted for by different intron lengths. This strongly suggests that the process of non-homologous recombination within introns required to duplicate the 54 bp exon and its derivatives has been active since the divergence of modern lineages from the ancestral fibrillar gene. Some authors have proposed that the dispersion of exons is a strategy to minimize the effects of non-homologous recombination by reducing the lengths of contiguous repeats available for misalignment during meiosis and by reducing the probability of a recombination event altering mRNA and hence peptide length by deleting or inserting exons. The only change that has been tolerated, the fusion of exons in COL1A1, does not change peptide length. Whether or not this is an evolved strategy, there are persuasive arguments that the evident suppression of such recombination events is essentially a functional one. Some connective-tissue disease mutants have one or more exons deleted from the helical domain (Chu *et al.* 1983; Sykes and Smith 1985). Because exons encode whole triplets, deletions or insertions do not disturb helix folding which proceeds normally from the C-terminus to the N-terminus because the glycine residues remain in register. The effect would be felt at the N-terminus of the helical region in molecules formed between deleted (or inserted) and full-length peptides. One apparent result is a reduced efficiency of N-propeptide cleavage which then interferes with molecular aggregation during and after fibril formation. The precise lateral alignment of lysines and hydroxylysines, essential for cross-linking, would also be disturbed. For much the same reason, a constraint on length variation would operate to prevent the independent evolution of genes whose products co-operate either in the formation of heteropolymeric molecules like collagens 1 and 5 or of mixed fibrils constructed from more than one collagen. As mentioned above, both collagens 3 and 5 may act in this way.

If this argument is correct and the strong inhibition of helix length change is a functional consequence of molecular and supramolecular construction, then the process of helix elongation to the standard length simply could not have occurred in a fibrillar collagen gene even before divergence of the modern lineages.

A HYPOTHETICAL SOLUTION

One radical solution to the paradox would see the major helical domain evolving elsewhere in the genome before insertion *en bloc* into an acceptor gene encoding the N- and C-terminal propeptides in a dramatic example of exon-shuffling. What follows is entirely speculation.

The independent evolution of the triplet domain might have occurred in either expressed or unexpressed sequences elsewhere. Families of tandemly repeated oligonucleotides occur frequently throughout the genome and appear to have evolved by processes unrelated to mutation and selection which normally govern expressed genes (Dover 1980). Repeating units can multiply rapidly and gene conversion can act to homogenize nucleotide sequence. Is it possible that the 54 bp exon and flanking splice sites might have formed part of a repetitive element which underwent rapid expansion? The mechanism would, in any event, be the same as we have already considered for the growth of the region in an expressed sequence, i.e. non-homologous recombination.

Perhaps the initial 54 bp exon originally formed part of a functional gene which had been inactivated by mutation. Removed from the functional constraints to helix growth, the original repeat and its simple derivatives could have duplicated to the present length, or longer, before reinsertion into an active acceptor gene. What sort of gene will be considered later. Nevertheless, if this argument is correct, the recombination event which incorporated the full length helix into an active gene would have been of immense significance. At a stroke, it would now be possible to construct a fibrillar collagen molecule. Instantly a new building material became available which eventually came to dominate vertebrate construction. The exon distribution was frozen by this new function and has remained unchanged ever since. It is hard to see how this model can be tested but consideration of the structure of the gene segments encoding the N- and C-propeptides is not against it.

THE N-PROPEPTIDE

Sequences cleaved from the N-terminus of fibrillar procollagens are encoded in seven exons (Fig. 1.4). From the amino acid sequence four distinct domains are distinguishable. Exon 1 contains the 5′ untranslated region in which the ATG start codon is preceded by two identical codons each followed by short open-reading frames. There is no evidence that these short peptides are expressed. The start codon is contained within a conserved sequence with the potential to form a cruciform structure. The 22-residue signal peptides and the signal peptidase site are both encoded by exon 1. The next domain, a cysteine-rich globular region, shows considerable variation between different genes. The human COL1A1 locus encodes an 86-residue domain distributed between the 3′ end of exon 1, the whole of exon 2 and the 5′ end of exon 3. The equivalent domains in mouse and chick lack 9 and 7 residues, respectively. However, it is in the comparison between COL1A1 and COL1A2 that the difference is most striking. Though still distributed between the same three exons, the sequences at COL1A2 encode a vestigial domain of only 8 residues. COL1A2 exon 2 is only 11 bp compared with 195 bp at COL1A1 and has the unusual sequence

cag ATGTGAGTGAG gtcagtatgatta

which, with the 3′ flanking sequence, consists of four overlapping consensus donor splice junctions found at the 5′ end of introns. There is no evidence

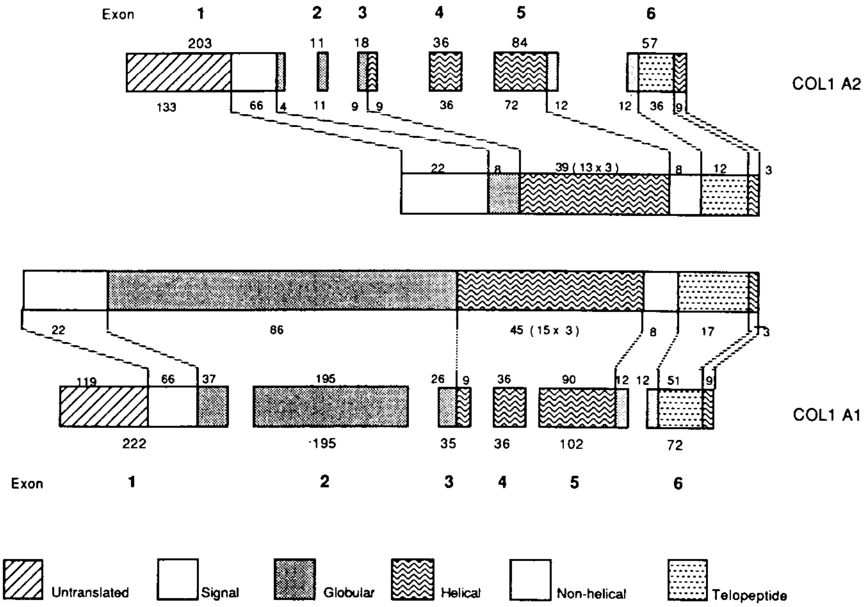


Fig. 1.4 The N-terminal domains of COL1A1 and COL1A2. Figures for amino acid and nucleotide lengths are taken from chick (COL1A2) and human (COL1A1) sequences.

that any other than the correctly positioned sequence is used. If they are, then the transcripts are unstable because they cannot be detected by S1 protection. The most straightforward explanation is that the sequences encoding the bulk of the domain have been lost from COL1A2 exon 2. The genes encoding collagen 2 (COL2A1) and collagen 3 (COL3A1) also show substantial differences in the structure of this region when compared to COL1A1. While relatively short sequences are missing from COL3A1 approximately two-thirds of the region is absent from COL2A1.

The function of the region is unknown. The conspicuous lack of conservation between genes which are otherwise remarkably similar suggests either that this domain is no longer essential or, alternatively, that it has evolved differently in different genes to modify their tissue-specific expression. There is some controversial evidence that the N-terminal telopeptide might act as a feedback inhibitor of collagen synthesis though this activity has not been specifically located in the globular region (Paglia *et al.* 1981). Unlike the highly-conserved C-terminal propeptide, which initiates triple-helix folding from the C-terminus via the formation of interchain disulphide bonds, the N-terminal propeptide is not required for folding of the major helix. The disulphide bonds are intra- rather than inter-chain. It could be that this region is a redundant vestige of a formerly functional domain.

Unlike the globular region, the N-terminal helical domain has been conserved between loci and between species. The Gly–X–Y triplet sequences are encoded by the 3′ end of exon 3, the whole of exon 4, and most of exon 5. The length of the Gly–X–Y stretch varies between COL1A1 (15 triplets) and COL1A2 (13 triplets) so clearly the maximum length of the triple helical domain must be 13 triplets since all three chains—two from COL1A1 and one from COL1A2—are needed to form a helix. The most interesting feature of the sequences encoding the triplet sequences is that they do not appear to be closely related to those encoding the major helix. Exon 4, the only one encoding entirely triplet sequences, is 36 bp which is not a size found in the major helix exons. Further, codon usage can be significantly different. In the chick COL1A2 gene, for instance, the third-position preference for proline switches from T (0.70 in the major helix, 0.29 in the minor helix) to A (0.20 in the major helix, 0.64 in the minor helix). Of other collagen genes, only those from *C. elegans* and *S. purpuratus* show a marked preference for the CCA proline codon. Similarly, the GGG glycine codon occurs twice in 13 opportunities in the minor helix (0.15) and only six times in 338 opportunities (0.02) in the major helix. Whether or not these codon-usage differences between major and minor helices are significant as an indication of their separate origins is not clear especially so when codon usage for the major helical domains of COL1A1 and COL1A2 are also very different.

Nevertheless, there are similarities in the overall organization of the sequences encoding the minor and major helices. Both are flanked at either end by junction exons, that is, by exons encoding both triplet and non-triplet sequences. Exons devoted to triplet coding encode only integral numbers of triplets beginning with the Gly and ending with Y positions of the Gly–X–Y repeat and, unlike some other collagen genes, the glycine codons are not split by introns. The 3′ end of exon 5 encodes four amino acids of an eight residue non-helical stretch which precedes the N-protease cleavage site. The remaining four residues, the cleavage site itself, the N-telopeptide and the first triplet of the major helix are all encoded in exon 6.

THE CARBOXY PROPEPTIDE

Sequences cleaved from the C-terminus of fibrillar procollagens are encoded in four exons (Fig. 1.5). Since major helix formation proceeds from the carboxy-terminus an inferred function of the propeptide is to bring the three chains into the correct juxtaposition for this to happen. Certainly, chains with mutations in the propeptide cannot be incorporated into triple helical molecules (Pihljaniemi *et al.* 1984). Another interesting activity has recently been discovered for the propeptide of collagen 2 after release by the C-protease. It appears to be identical to chondrocalcin which recognizes a receptor on chondrocyte cell surfaces and promotes cell binding to collagen (Van der Rest *et al.* 1986). Whether similar receptors exist for the C-propeptides of other collagens remains to be seen.

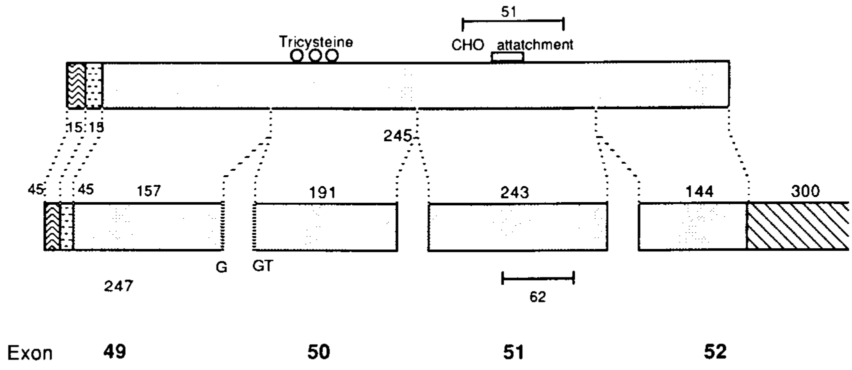


Fig. 1.5 The C-terminal domains of chick COL1A2.

As at the N-terminus of the major helix, the C-terminal triplets, telopeptide, and protease cleavage site are encoded in a single junction exon. Because of the effect on exon numbering of the fusion of two 54 bp exons in the major helix in COL1A1, the junction exon is either 48 (COL1A1) or 49 (COL1A2). To avoid confusion and, as an acknowledgement of sequencing precedence, the COL1A2 numbering will be used.

Exon 49 encodes the five C-terminal triplets of the major helix, four of which are contiguous Gly–Pro–Pro repeats, also a feature of the N-terminal triplets of the minor helix in the N-propeptide but nowhere else in either helix. It is an attractive suggestion that these 'ideal' triplets occur at the carboxy ends of helices to get the folding off to a good start. In COL3A1, this exon encodes an additional two triplets depending on species making the major helix length 340 rather than the 338 triplets in COL1A1, COL1A2 and COL2A1. This apparently minor difference is nevertheless the only example of divergence in the structure of the major helix of any fibrillar collagen and it is interesting that the departure from the strict rules of structure has only been allowed in a junction exon. There seems to be no rational explanation for this.

The 245-residue (COL1A2, COL3A1) or 246-residue (COL1A1, COL2A1) C-propeptide is encoded by the 3′ end of exon 49 and exons 50–52. Some authors recognize distinct 'subdomains' encoded by each separate exon though without a more detailed knowledge of propeptide functions and tertiary structure this is an unsupported argument. Exons 50 and 51 encode the five cysteine residues which contribute to the inter-chain disulphide bonds which stabilise the three chains of the propeptide. Exon 51, in addition, encodes a site for the addition of an N-linked heteropolysaccharide chain. Finally, exon 52 is both the largest exon and the only one to vary in length. The 5′-most 144 bp encode 38 residues before the termination codon and at least 300 bp of untranslated transcript. The multiple mRNA transcripts which have been demonstrated for all fibrillar genes result from a