THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 1988 by The American Society for Biochemistry and Molecular Biology, Inc.

Vol. 263, No. 34, Issue of December 5, pp. 18328-18337, 1988
Printed in U.S.A.

Drosophila Basement Membrane Procollagen $\alpha 1$(IV)

II. COMPLETE cDNA SEQUENCE, GENOMIC STRUCTURE, AND GENERAL IMPLICATIONS FOR SUPRAMOLECULAR ASSEMBLIES*

(Received for publication, March 8, 1988)

Bruce Blumberg‡§, Albert J. MacKrell¶‖, and John H. Fessler¶‡**

From the ¶Molecular Biology Institute and ‡Department of Biology, University of California at Los Angeles, Los Angeles, California 90024

A *Drosophila melanogaster* gene for a basement membrane procollagen chain was recently identified from the sequence homology of the carboxyl (NCl) end of the polypeptide that it encodes with the corresponding domain of human and murine collagens IV (Blumberg, B., MacKrell, A. J., Olson, P. F., Kurkinen, M., Monson, J. M., Natzle, J. E., and Fessler, J. H. (1987) *J. Biol. Chem.* 262, 5947–5950). This gene is at chromosome location 25C. Here we report the complete 6-kilobase cDNA sequence coding for a chain of 1775 amino acids, as well as the genomic structure. The gene is composed of nine relatively large exons separated by eight relatively small introns. This organization is different from the multiple small exons separated by large introns reported for mouse and human type IV collagens (Kurkinen, M., Bernard, M. P., Barlow, D. P., and Chow, L. T. (1985) *Nature* 317, 177–179. Sakurai, Y., Sullivan, M., and Yamada, Y. (1986) *J. Biol. Chem.* 261, 6654–6657. Soininen, R., Tikka, L., Chow, L., Pihlajaniemi, T., Kurkinen, M., Prockop, D. J., Boyd, C. D., and Tryggvason, K. (1986) *Proc. Natl. Acad. Sci. U. S. A.* 83, 1568–1572).

***Drosophila* and human $\alpha 1$(IV) procollagen chains share not only polypeptide domains near their amino and carboxyl ends for making specialized, intermolecular junctional complexes, but also 11 of 21 sites of imperfections of the collagen triple helix. However, neither the number nor the nature of the amino acids in these imperfections appear to have been conserved. These imperfections of the helical sequence may be important for the supramolecular assembly of basement membrane collagen.**

The 9 cysteine residues of the *Drosophila* collagen thread domain are arranged as several variations of a motif found in vertebrate collagens IV only near their amino ends, in their "7 S" junctional domains. The relative positions of these cysteine residues provide numerous opportunities for disulfide bonding between molecules in both parallel and antiparallel arrays. There is a pseudorepeat of one-third of the thread length, and there are numerous possibilities for disulfide-linked microfibrils and networks. We propose that collagen microfibrils, stabilized by disulfide segment junctions, are a versatile ancestral form from which specialized collagen fibers and networks arose.

Vertebrate type IV collagen is a triple-helical molecule which can exist in both heterotrimeric (5) or homotrimeric forms (6). In the electron microscope the molecule appears as a thread of approximately 400 nm with a prominent knob at the carboxyl end (7). The thread domain corresponds mostly to collagen triple helix, but there are imperfections within the (Gly-X-Y)$_n$ sequence that is required for the collagen fold. Correspondingly, the flexibility of this thread is not uniform (8). Each collagen IV chain extends as a noncollagenous sequence, called NCl (9), to form the carboxyl knob. The complete primary structures of human $\alpha 1$(IV) (10) and mouse $\alpha 2$(IV)[1] have recently been determined. Individual triple-helices can associate via their amino termini to form tetramers (11) and via their carboxyl termini to form dimers (12). In addition to these end-to-end interactions, extensive lateral associations of type IV collagen molecules have been deduced (13). Several models have been proposed for the structure of type IV collagen networks based both on observations of *in vitro* associations among molecules (11, 14) and on the structure of acid-extracted material from lens capsule basement membranes (13). There are data which support each model and although the manner of organization differs in the different models, it is clear that type IV collagen is organized into networks which may provide the scaffold around which basement membranes are built. Because the structure of basement membranes is likely to vary depending on the tissue in which it is found (13), it is possible that no single model is sufficient to account for the full variety of basement membranes.

We are analyzing the assembly, structure, and function of basement membranes in *Drosophila melanogaster*, which allows genetic, biochemical, and recombinant DNA manipulations. Previously, we reported that a collagen, laminin, entactin, and a proteoglycan are produced by cultures of *Drosophila* Kc cells (15). Detailed biochemical characterizations of the laminin (16), proteoglycan (17 and 18), collagen (19, 50), and entactin[2] are proceeding. It is clear from these analyses that several *Drosophila* basement membrane components are biochemically very similar to their vertebrate counterparts. In accord with the biochemical data, we have previously reported that portion of the sequence of a cDNA which codes for about 40% of the pro$\alpha 1$(IV) polypeptide of *Drosophila*, including the carboxyl end. The carboxyl-terminal 231-residue sequence is

* This work was supported in part by United States Public Health Service Grant AG-02128. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ Supported by National Research Service Award CA-09056. Present address: Dept. of Medicine, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, Piscataway, NJ 08854.

‖ Supported by National Research Service Award GM-07185.

** To whom correspondence should be addressed.

[1] M. Kurkinen, S. Quinones, J. Saus, A. J. MacKrell, and B. Blumberg, manuscript in preparation.

[2] P. F. Olson, L. I. Fessler, R. Sterne, R. Nelson, A. G. Campbell, and J. H. Fessler, manuscript in preparation.

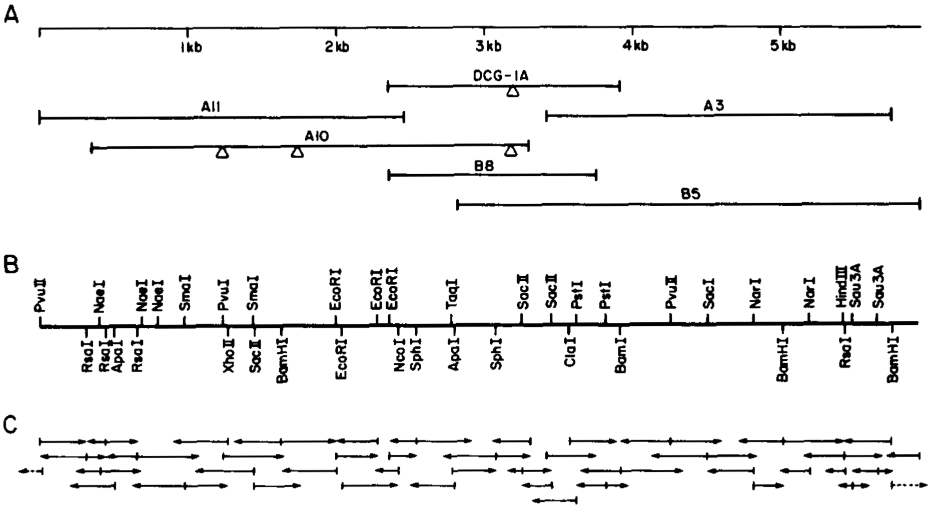

FIG. 1. ***Drosophila* procollagen cDNA clones.** *A*, overlapping cDNA clones CDA10, CDA11, CDA3, CDB5, and CDB8 are shown. Clone CDA10 fortuitously contains introns and these are shown as *triangles*. The cloned genomic DNA fragment, DCG-1A, which was used to screen the cDNA library, is shown for reference. *B*, partial restriction map of the composite cDNA, showing sites used for subcloning restriction fragments for DNA sequencing. *C*, restriction fragments with termini shown by *vertical lines* were subcloned and sequenced in the direction indicated by the *arrows*. *Broken lines* indicate fragments obtained from genomic clones.

highly homologous with the corresponding domain of human and murine collagens IV (1). This gene is at chromosome location 25C. Here we report the complete sequence of *Drosophila* proα1(IV), as deduced from cDNA, and its complete genomic organization. We compare the sequence of this *Drosophila* collagen with those of the human α1(IV) and mouse α2 (IV) chains[3] and then discuss the implications of our results for a general understanding of the collagen scaffold of basement membranes.

MATERIALS AND METHODS[4]

RESULTS

We used the previously described *Drosophila* genomic DNA clone DCG1A (20) to screen a cDNA library prepared from a *Drosophila* cell line (1). We isolated and purified a number of clones which together cover the entire coding sequence of *Drosophila* proα1(IV). The relationship of these clones to each other and to DCG1A is shown in Fig. 1*A*. A composite restriction map of these overlapping clones is shown in Fig. 1*B*. We sequenced overlapping restriction fragments from the various cDNA clones as described (1). The sequencing strategy is shown in Fig. 1*C*.

The complete coding sequence deduced from the overlapping cDNA clones is shown in Fig. 2. The sequence was completely determined on both strands and across most restriction sites used in cloning. Each nucleotide is represented, on average, 7.5 times in the sequencing database. The cDNA sequence from nucleotides 3418–5946 has been previously reported by us (1) and is presented here for completeness and to facilitate discussion of the sequence as a whole. The sequence encodes 1775 amino acids including a putative signal peptide. This is somewhat larger than the 1669-residue size of human α1(IV) (10) but is consistent with the observation that the *Drosophila* procollagen we isolated from cell cultures is approximately 30 nm longer than mouse collagen IV (7) which would correspond to approximately 100 more amino acids assuming a triple helical length of 0.29 nm/residue (21).

Genomic Organization—We prepared ^{32}P-labeled RNA probes as described (1) from both the 5′ and 3′ ends of the cDNA and used them to screen a *Drosophila* genomic DNA library (22) at high stringency. We isolated and mapped 10 independent overlapping clones spanning approximately 35 kb[5] of genomic DNA. Maps of these clones are shown in Fig. 3. A detailed restriction map was obtained for the genomic clones using enzymes which recognize sites in the cDNA, and introns were located by comparing this map with that for the cDNA. Appropriate restriction fragments were obtained from genomic clones and sequenced to establish their precise boundaries and sizes of the introns. The coding region of the gene maps to a region of about 7.5 kb and is indicated by a *bold line* in Fig. 3. Since the coding region is flanked by at least 13 kb of genomic DNA in our clones it is likely that we have isolated the complete gene including any 5′ or 3′ regulatory elements. However, this remains to be demonstrated. The 8.2-kb *Eco*RI fragment corresponds to the clone DCGA1 (20), which was isolated from the same library, as judged by partial DNA sequence analysis and restriction mapping.

Intron/Exon Structure—A schematic drawing of the intron/exon structure of *Drosophila* proα1(IV) is shown in Fig. 4. We number the introns and exons from the 5′ end of the gene. The gene consists of 9 exons of 172, 101, 49, 948, 497, 1452, 979, 1132, and 638 bp separated by introns of 335, 484, 64, 75, 125, 62, 90, and 269 bp. These sizes for the exons show no relationship to the multiple small exons demonstrated for mouse (2, 3) and human type IV collagens (4) and the major interstitial collagens (reviewed in Ref. 23). Furthermore, none of the exons are multiples of 9 bp or show any relationship to the postulated 54-bp primordial collagen exon (24). None of the published exon sizes for mouse α1(IV) (3) or mouse α2(IV) (2) bear any relationship to the 54-bp primordial exon, and the two chains of mouse type IV collagen are considered to have evolved differently from those for interstitial collagens (2, 3). The intron/exon boundaries of this *Drosophila* gene are given in Fig. 5 in the Miniprint.

Potential Cell-binding Domain—Both the *Drosophila* proα1(IV) and the human α1(IV) chains contain the sequence Arg-Gly-Asp-Thr (RGDT), which has been shown to be a functional cell-binding domain in vertebrate type I collagen (25). The positions in the *Drosophila* and human collagen IV

[3] The human α1(IV) and mouse α2(IV) sequences were kindly communicated to us, before publication, by, respectively, Dr. K. Kühn (Max-Planck Institute for Biochemistry, Martinsried) (10), and Dr. M. Kurkinen[1] (University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, Piscataway, NJ).

[4] Portions of this paper (including "Materials and Methods," part of "Results," Figs. 5, 8, and 10, and Tables 1–3) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are included in the microfilm edition of the Journal that is available from Waverly Press.

[5] The abbreviations used are: kb, kilobases; bp, base pairs.

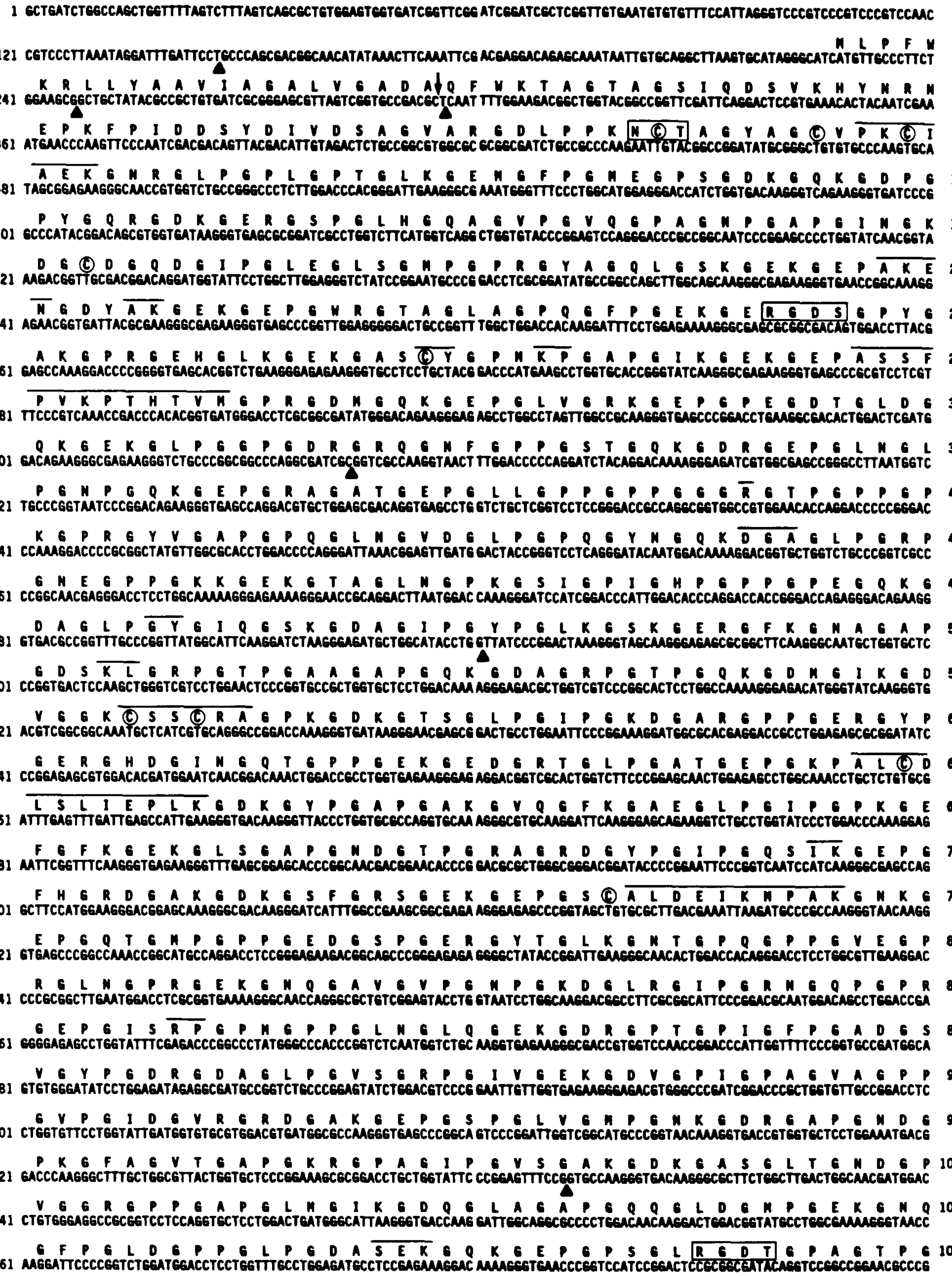

FIG. 2. **The complete nucleotide sequence and deduced amino acid sequence of the *Drosophila* proα1(IV) composite cDNA.** The sequence of the composite cDNA was determined using the strategy shown in Fig. 1*C*. The cDNA was sequenced entirely on both strands and across all restriction sites used in subcloning. The putative signal cleavage site and the boundary between the helical region and the NC1 are shown with *vertical arrows.* Imperfections in the collagen triple helix are indicated with *horizontal lines* and cysteine residues are *circled.* Potential cell-binding sites, RGDS and RGDT, and a potential *N*-glycosylation site, are *boxed.* The location of introns is indicated by *triangles* beneath the nucleotide sequence, and the consensus polyadenylated signal is *boxed.*

```
          W  P  G  E  K  G  L  P  G  L  A  V  H  G  R  A  G  P  P  G  E  K  G  D  Q  G  R  S  G  I  D  G  R  D  G  I  N  G  E  K  1125
3481 GTTGGCCAGGAGAGAAGGGTTTGCCCGGTCTGGCTGTTCACGGTCGTGCTGGTCCGCCAG GCGAGAAGGGTGACCAGGGACGCAGTGGAATCGATGGACGAGATGGAATTAACGGCGAGA

          G  E  Q  G  L  Q  G  V  W  G  Q  P  G  E  K  G  S  V  G  A  P  G  I  P  G  A  P  G  M  D  G  L  P  G  A  A  G  A  P  G  1165
3601 AGGGTGAACAAGGTCTGCAGGGCGTTTGGGGCCAGCCTGGCGAGAAGGGATCTGTCGGCG CACCTGGTATTCCTGGTGCTCCCGGAATGGATGGTTTGCCCGGCGCTGCTGGTGCTCCTG

          A  V  G  Y  P  G  D  R  G  D  K  G  E  P  G  L  S  G  L  P  G  L  K  G  E  T  G  P  V  G  L  Q  G  F  T  G  A  P  G  P  1205
3721 GTGCTGTTGGCTATCCTGGTGATCGCGGTGACAAGGGAGAGCCTGGTCTATCTGGTCTGC CCGGACTCAAGGGTGAGACTGGACCCGTTGGACTGCAGGGCTTCACCGGTGCTCCTGGCC

          K  G  E  R  G  I  R  G  Q  P  G  L  P  A  T  V  P  D  I  R  G  D  K  G  S  Q  G  E  R  G  Y  T  G  E  K  G  E  Q  G  E  1245
3841 CTAAGGGTGAGCGCGGTATTCGTGGTCAGCCCGGTCTTCCGGCCACCGTTCCCGACATTC GTGGTGATAAGGGATCCCAGGGCGAGCGCGGCTACACTGGCGAGAAGGGCGAGCAAGGCG

          R  G  L  T  G  P  A  G  V  A  G  A  K  G  D  R  G  L  Q  G  P  P  G  A  S  G  L  N  G  I  P  G  A  K  G  D  I  G  P  R  1285
3961 AACGAGGCTTGACTGGTCCTGCTGGCGTCGCTGGAGCAAAGGGAGATCGCGGATTGCAGG GACCACCAGGTGCAAGCGGATTGAACGGCATTCCCGGAGCCAAGGGAGACATTGGTCCAA

          G  E  I  G  Y  P  G  V  T  I  K  G  E  K  G  L  P  G  R  P  G  R  N  G  R  Q  G  L  I  G  A  P  G  L  I  G  E  R  G  L  1325
4081 GAGGCGAGATCGGTTATCCAGGAGTTACCATTAAGGGCGAGAAGGGTCTGCCCGGTCGCC CAGGCAGAAACGGACGTCAAGGTCTTATTGGAGCACCCGGCTTAATTGGAGAACGTGGTC
                                                                                        ▲
          P  G  L  P  E  S  R  L  V  G  L  P  G  P  I  G  P  A  G  S  K  G  E  R  G  L  A  G  S  P  G  Q  P  G  Q  D  G  F  P  G  1365
4201 TGCCTGGCTTGCCGGAGAGCCGCCTCGTGGGCCTGCCTGGACCCATTGGACCAGCTGGCA GCAAGGGAGAGCGTGGTCTCGCCGGCAGTCCCGGACAACCAGGACAGGATGGCTTCCCCG

          A  P  G  L  K  G  D  T  G  P  Q  G  F  K  G  E  R  G  L  N  G  F  E  G  Q  K  G  D  K  G  D  R  G  L  Q  G  P  S  G  L  1405
4321 GCGCACCTGGATTGAAGGGAGATACTGGACCGCAGGGCTTTAAGGGCGAACGTGGTCTGA ATGGCTTCGAGGGACAAAAGGGAGACAAGGGTGACCGAGGACTCCAAGGACCGTCTGGAC

          P  G  L  V  G  Q  K  G  D  T  G  Y  P  G  L  N  G  N  D  G  P  V  G  A  P  G  E  R  G  F  T  G  P  K  G  R  D  G  R  D  1445
4441 TGCCCGGCTTGGTTGGACAGAAGGGAGATACCGGCTACCCTGGCTTAAATGGAAACGATG GACCTGTCGGAGCTCCTGGCGAGCGCGGCTTCACCGGGCCCAAGGGACGCGATGGACGCG

          G  T  P  G  L  P  G  Q  K  G  E  P  G  M  L  P  P  P  G  P  K  G  E  P  G  Q  P  G  R  N  G  P  K  G  E  P  G  R  P  G  1485
4561 ACGGAACACCAGGTCTGCCTGGACAGAAGGGTGAACCAGGAATGCTGCCACCACCAGGAC CCAAGGGCGAACCTGGTCAGCCGGGACGCAATGGACCTAAGGGAGAGCCCGGACGTCCGG

          E  R  G  L  I  G  I  Q  G  E  L  G  E  K  G  E  R  G  L  I  G  E  T  G  N  V  G  R  P  G  P  K  G  D  R  G  E  P  G  E  1525
4681 GAGAGCGTGGCTTGATTGGCATCCAGGGTGAGCTTGGCGAAAAGGGTGAGCGCGGCCTGA TCGGTGAGACTGGTAACGTGGGACGACCCGGACCCAAGGGAGATCGCGGAGAGCCAGGCG

          R  G  Y  E  G  A  I  G  L  I  G  Q  K  G  E  P  G  A  P↓ A  P  A  A  L  D  Y  L  T  G  I  L  I  T  R  H  S  Q  S  E  T  1565
4801 AGAGGGGATATGAGGGCGCCATTGGTTTGATCGGCCAGAAGGGTGAGCCCGGTGCTCCAG CCCCCGCTGCTCTGGACTATCTCACCGGTATCCTGATTACGCGACACAGTCAATCGGAAA

          V  P  A  Ⓒ  S  A  G  H  T  E  L  W  T  G  Y  S  L  L  Y  V  D  G  N  D  Y  A  H  N  Q  D  L  G  S  Ⓒ  V  P  R  F  S  T  1605
4921 CGGTGCCCGCTTGCTCGGCTGGACACACGGAACTGTGGACGGGTTACTCCCTGTTGTACG TCGATGGCAATGACTATGCCCACAACCAGGACCTTGGATCCTGTGTGCCACGCTTCTCAA

          L  P  V  L  S  Ⓒ  G  Q  N  N  V  Ⓒ  N  Y  A  S  R  N  D  K  T  F  W  L  T  T  N  A  A  I  P  M  M  P  V  E  N  I  E  I  1645
5041 CGCTGCCCGTACTGTCGTGTGGTCAGAATAACGTCTGCAACTACGCCTCCAGAAATGACA AGACCTTCTGGCTGACAACCAACGCCGCCATTCCGATGATGCCCGTTGAAAACATCGAGA

          R  Q  Y  I  S  R  Ⓒ  V  V  Ⓒ  E  A  P  A  N  V  I  A  V  H  S  Q  T  I  E  V  P  D  Ⓒ  P  N  G  W  E  G  L  W  I  G  Y  1685
5161 TCCGCCAGTACATCTCACGTTGCGTCGTTTGTGAGGCGCCGGCTAATGTGATCGCCGTGC ACAGTCAAACGATAGAGGTGCCCGACTGTCCGAATGGCTGGGAGGGTCTCTGGATTGGCT

          S  F  L  M  H  T  A  V  G  N  G  G  G  G  Q  A  L  Q  S  P  G  S  Ⓒ  L  E  D  F  R  A  T  P  F  I  E  Ⓒ  N  G  A  K  G  1725
5281 ACAGTTTCCTCATGCACACTGCCGTGGGCAACGGTGGCGGTGGACAGGCGCTGCAATCGC CTGGCTCCTGTTTGGAGGACTTCCGTGCAACGCCCTTCATCGAGTGCAACGGCGCCAAGG
                   ▲
          T  Ⓒ  H  F  Y  E  T  M  T  S  F  W  M  Y  N  L  E  S  S  Q  P  F  E  R  P  Q  Q  Q  T  I  K  A  G  E  R  Q  S  H  V  S  1765
5401 GCACGTGCCACTTCTATGAGACGATGACAAGCTTCTGGATGTACAACCTGGAGTCCTCAC AGCCGTTCGAGAGGCCACAGCAGCAGACGATCAAGGCCGGCGAGCGGCAGTCGCATGTGT

          R  Ⓒ  Q  V  Ⓒ  M  K  N  S  S  *                                                                                        1775
5521 CCAGGTGCCAGGTGTGCATGAAGAACTCCTCGTAGGACCTCCAATCCCAACACAGACACA CCCACAGCACAGAGCATAAGTTTAATCTAAATGTAAAGCCTTAAAATTACCAACGAATCG

5641 TGTGCACACCCACACGATCACAACACAAACACAAACAAACAAACCAACACACACACACAC ATACATACACCGACGAACCAACACTGCTACAAATTCCTTAATCTAACCAAAAAAAAAAAA

5761 AAACGGATCCACAAGTCGAAGTGCTAATTACCGACCGACCTGGACCACAATGCCATTTTT TATCTGCCAATTAATGTTCTAAACAAATGTAAACTGCTTCTAAGTTACGTTACGTATGTT

5881 AAGTAAATGAAACAAATAAATAAACTAGCCTCGGTCGTATGTCCAAAAGTATAAAAAAAA AAAAAA
```

FIG. 2—*continued*

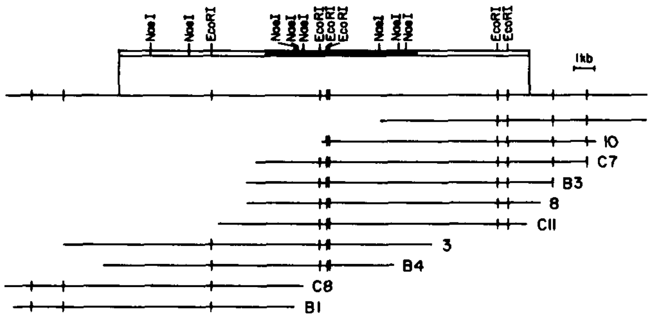

FIG. 3. **Genomic clones for *Drosophila* proα1(IV).** Ten independent, overlapping clones covering the gene for *Drosophila* proα1(IV) were isolated, and their *Eco*RI restriction maps are shown. The *top line* is a composite restriction map derived from clones B4 and C11. The mRNA coding region is 7.5 kb in length, and is indicated by the *black bar.*

chains are, respectively, 466 amino acids and 840 amino acids from the NCl domain. The *Drosophila* proα1(IV) also contains the sequence Arg-Gly-Asp-Ser (RGDS) 1307 amino acids from the NCl domain. RGDS has been shown to be a cell receptor binding domain in fibronectin and other molecules (26). Either of these regions could be sites where type IV collagen is bound by cell-surface receptors.

Imperfections of Triple Helix—The sequence shown in Fig. 2 contains 22 imperfections of the (Gly-*X*-*Y*)$_n$ repeat necessary for a collagen triple helix. We count as an imperfection the set of amino acids that follow the *Y* of a Gly-*X*-*Y* repeat and precede the next Gly-*X*-*Y* triplet. As the first two imperfections are only separated by one (Gly-*X*-*Y*), we consider the pair as a single, fused imperfection from here on. The imperfections are listed in Table I and shown in Fig. 6. The position of each imperfection is denoted by the residue at its amino end in two ways. First, as its sequential number from the amino end of the unprocessed polypeptide. Second, to help comparison with vertebrate collagen IV chains, the junction between the collagen thread and the carboxyl (NCl) domain is taken as origin, and residues are numbered sequentially toward the amino end of the polypeptide. Two-thirds (14:21) of the imperfections of the *Drosophila* proα1(IV) chain correspond in their locations to imperfections in either the human α1(IV) collagen chain or in the mouse α2(IV) collagen chain. The positions of more than one-third of the *Drosophila* proα1(IV) chain imperfections occur in both vertebrate

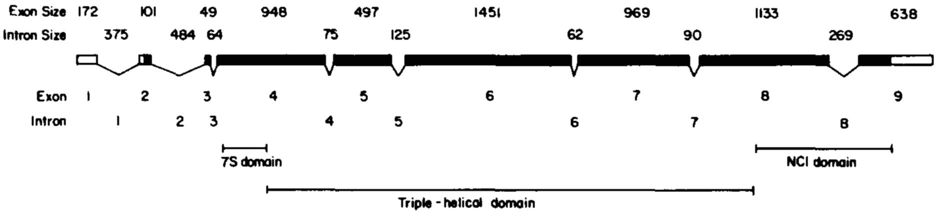

FIG. 4. **Intron/exon structure of the *Drosophila* proα1(IV) gene.** A schematic representation of the mRNA encoding region of *Drosophila* proα1(IV) gene is shown with exons represented by *boxes* and introns by *lines*. The protein-coding regions are *shaded*. Exons and introns are numbered from the 5′-end and their sizes are indicated above their locations in the diagram.

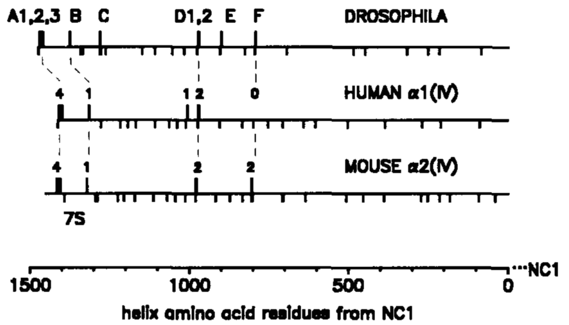

FIG. 6. **Relative locations of cysteine residues and imperfections of helix in *Drosophila* proα1(IV), human α1(IV), and mouse α2(IV) chains.** The collagen thread portions of the three chains are diagrammed as *horizontal lines*, aligned at their thread/NCl domain junctions. Each junction is taken as origin for the coordinate of residue numbers. The positions of cysteine residues are indicated by the *longer vertical bars*, above each horizontal line. The alphanumeric designation of each thread cysteine residue of *Drosophila* proα1(IV) is shown. The small numbers associated with the human α1(IV) and mouse α2(IV) chains give the number of cysteine residues in clusters that cannot be resolved at the scale of this diagram (or the lack of a cysteine residue). *Dashed lines* indicate suggested correspondence of locations of cysteine residues between different species. The 7 S region of the vertebrate chains is indicated. The 2 cysteine residues indicated for mouse α2(IV) at approximately coordinate 800 are taken to form a disulfide link. The corresponding loop-out of the polypeptide between them is not shown. This is about 20 amino acids long and is allowed for in the placement of all residues of this mouse chain at the amino end. Imperfections of helical sequence are indicated by *short vertical lines* below the *horizontal lines*.

chains. Fig. 7 illustrates the matching of 11 imperfections of the *Drosophila* proα1(IV) chain to those of the human α1(IV) chain. Note that while the origin of each axis corresponds to the carboxyl end of that collagen thread, the linear relationship of this comparison neither necessitates that the regression line should pass through this origin, nor that the positions of imperfections are identical, but only requires proportional displacements of imperfections in the pair of chains. On average each of these imperfections of the *Drosophila* proα1(IV) chain is within 4 residues of that predicted by the regression line and none are further away than 8 residues. While this regards the 22 imperfections as point residues, their actual average length is 5 ± 3 residues. Therefore the error of prediction approximately equals the variability of the length of the imperfections, which is not correlated in the two species. The prediction of position is relative to the positions of the other imperfections and is not absolute. As the actual thread lengths of the *Drosophila* and human chains are different, this correlation suggests that either the differences are primarily at the ends of the threads, and/or additional amino acids are relatively uniformly distributed throughout the thread. Similar comparisons were made with the mouse α2(IV) collagen chain and between the two vertebrate chains (not shown). All but one of the 14 *Drosophila* imperfections between position 1140 from NCl and the NCl/thread junction have equivalents in one or both vertebrate chains. The inverse does not hold: the human and mouse chains have several imperfections in this range which do not seem to have counterparts in *Drosophila.*

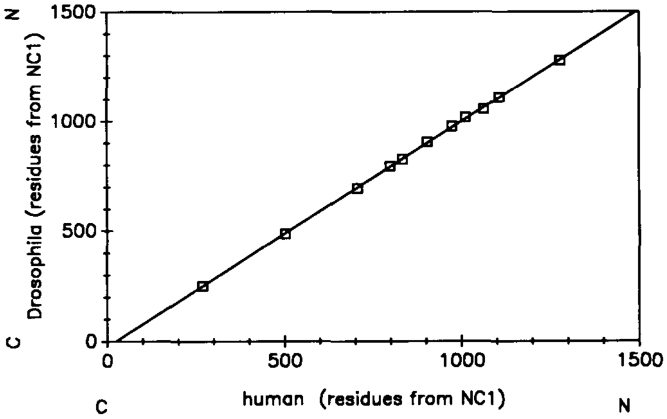

FIG. 7. **Alignment of some of the *Drosophila* and human collagen IV thread imperfections.** Initially an orthogonal grid was drawn, with *horizontal lines* denoting the positions of imperfections in the *Drosophila* proα1(IV) chain and *vertical lines* those of the human α1(IV) chain. For clarity, only the set of intersections of this grid which lie close to an approximately *diagonal line* are shown. The figure shows the regression line computed through these intercepts. The slope of the line is 1.022 ± 0.005, and the intercept on the *Y* (*Drosophila*) *axis* is −25 ± 5 residues. The carboxyl end of the collagen thread domain is indicated by *C*, and the amino end by *N*.

The amino acids within the imperfections account for about 6% of the residues in the *Drosophila* collagen thread. They are not conserved and can be charged, polar or hydrophobic. Cysteine occurs 10 times as frequently in imperfections than in collagen sequence and particularly occurs in longer imperfections. Glycine and proline occur less frequently in imperfections. No preference of amino acids in imperfections relative to collagen helix was discerned that was similar in the *Drosophila* proα1(IV), human α1(IV), and mouse α2(IV) collagen chains.

Cysteine Residues in the Collagen Thread Domain—The cysteine residues in the thread domain of this *Drosophila* chain are compared with those of the human α1(IV) and mouse α2(IV) collagen chains in Fig. 6 and are tabulated in Table II. Each of the 9 cysteine residues is referred to by an alphanumeric name, and they form two pseudohomologous sets, [A1,A2,A3,B,C] and [D1,D2,E,F]. Each set is the approximate equivalent of two vertebrate 7 S regions in tandem. The length of the 7 S region in the human α1(IV) collagen chain varies from 94 to 84 residues, depending on which of the 4 cysteine residues at its amino end are considered as a cysteine-cysteine interval (Fig. 6). In the *Drosophila* proα1(IV) chain there is a corresponding set of 3 cysteine residues [A1,A2,A3], and they are 95–84 residues from [B]. This is immediately followed by a similar, 95-residue interval [B–C].

Whereas cysteine [B] is at the center of symmetry of the interval [A1–C], the corresponding cysteine [E] is placed asymmetrically (2:3) in the interval [D1-E-F]. The two sets can be partly mapped onto each other by a transposition of 489 residues, which is one-third of the thread length. The pair of cysteine residues [D1,D2] have vertebrate equivalents in both the human α1(IV) and mouse α2(IV) chains. Cysteine residue [F] is matched by a pair of cysteine residues in the mouse α2 chain which reside in one large helix imperfection. Several lines of reasoning all show that this imperfection is looped out and presumably stabilized by a disulfide bond between the two cysteine residues which are thereby brought into apposition. Alignment of the other cysteine residues of the mouse α2(IV) collagen chain, and of its imperfections of Gly-*X*-*Y* sequence, with both the *Drosophila* proα1(IV) chain and the human α1(IV) collagen chain require this, and this is assumed in Fig. 6.

Computer alignments of the *Drosophila* collagen IV [A1-B-C] region with the human α1(IV) and mouse α2(IV) chains are, respectively, 14 and 16 standard deviation units from the means obtained by aligning these sequences after randomization. These alignments are shown in Fig. 8 and are much better than between other parts of the collagen thread regions of these chains. They are complexly influenced by both the Gly-*X*-*Y* sequence constraint and by helix imperfections of different lengths. The result is in agreement with our comparison of the distribution of cysteine residues in these chains (Fig. 6 and Table II). We suggest that not only are the elements of the vertebrate 7 S junction found in *Drosophila* collagen IV but also that the vertebrate chains may contain parts of the more extensive potential junctional regions of *Drosophila*.

DISCUSSION

In a previous paper (1) we described the extensive similarities between the carboxy-terminal (NC1) domains of *Drosophila* proα1(IV) and mouse and human α1(IV) and mouse α2(IV). These similarities were reflected in sequence homology at the DNA and amino acid level, in hydropathy profiles, and in three-dimensional structure as revealed by hydrophobic correlation coefficients. From this analysis we concluded that the junction between six carboxyl peptides is a fundamental molecular junction between collagen triple helices in basement membranes. The sequences at the amino termini of the molecules are also partly conserved. Two important cross-linking domains have been conserved through more than 500 million years of evolution, the approximate time at which the lines leading to vertebrates and invertebrates diverged (27). In contrast, the genomic organization has not been conserved between *Drosophila* and vertebrate collagens IV.

Exon-Intron Structure—The first exon and intron are in the 5′-untranslated part of the message. The second intron is in the hydrophobic part of the signal peptide, and the third intron is in a noncollagenous sequence, just preceding the first cysteine-rich region. In the DNA encoding the next 1453 amino acids, which comprise the collagen helix with its imperfections, there are only four introns: numbers 4, 5, 6, and 7. The eighth, and last intron occurs within a highly conserved amino acid sequence of the noncollagenous carboxyl peptide, and in precisely the same place as the penultimate 3′ intron in mouse α1(IV),[6] human α1(IV) (28), and mouse α2(IV).[6] Because only four of the eight introns are associated with the collagen helical domain, there is no preferential association of introns with the collagen helix. We conclude that there is no indication that introns are associated with maintenance of the Gly-*X*-*Y* genomic structure of this collagen. All exon-exon junctions which lie in this region are located at proper helix sequences and not at interruptions of it; therefore, introns are not associated with imperfections in the collagen triple helix either.

A curious repeat distance of about 484 residues arises in several ways. First, this is one-third of the 1453-residue sequence. Second, the cysteine residues found in this main portion of the molecule show a distinct pseudo-repeat separated by 489 residues (Table II). Third, the sixth exon, between introns 5 and 6, is 484 amino acids long. Last, the combined lengths of exons 4 and 5, spanning the sequence between introns 3 and 5, is 482 amino acids and is practically colinear with one of the above pseudo-repeats. This suggests that intron pairs 3 and 5, and 5 and 6, are the boundaries of a twice-repeated sequence of about 484 amino acids. A third repeat would predict an intron at the junction of the thread and NC1 domains, but this is missing. However, there are introns in this approximate location in human α1(IV) (4), mouse α1(IV) (3), and mouse α2(IV).[6]

Segment Junctions—We use the term *segment junction* for the general type of junction that is formed when two adjacent collagen threads become covalently linked to each other. By this definition, the vertebrate 7 S junction of four collagen IV molecules is a segment junction, and so is the 60-nm overlap of two antiparallel collagen VII molecules (29). A segment junction contains two or more collagen threads in parallel or antiparallel orientation, extends over only a fraction of the component collagen helices, and may include imperfections of helix. A segment junction is held by at least two covalent bonds, one near each end, and by any number of additional bonds.

Table III lists regions of the *Drosophila* proα1(IV) chain that are well suited for segment junctions stabilized by two or more disulfide and/or lysine-derived bonds. Each of these sequences in one homotrimeric molecule matches the same sequence in an adjacent, antiparallel collagen thread. The listed lysine residues are all pseudosymmetrically placed except in set 3. The hydropathy plot (30) of set 5 is fairly symmetrical about the centrally placed [B] cysteine (not shown). Approximately 30-nm long segment junctions will be formed by sets 1–4 and about 60-nm junctions by sets 5 and 6. Electron microscopy of *Drosophila* procollagen molecules has shown some oligomers with equivalent overlaps of the amino ends of the molecules (50). SDS-agarose gel electrophoretic analyses of the oligomers indicated a pronounced dimer, and then decreasing amounts of higher forms, without showing the predominant amount of tetramer found for vertebrate collagen IV (31, 50).

Microfibrils—Disulfide bonds could also form between different sequence regions of two *Drosophila* collagen molecules. The similar cysteine-rich regions [A1,A2,A3-B-C] and [D1,D2-E-F] of two molecules can be juxtaposed by a displacement of 489 amino acids, one-third of the collagen thread sequence (Table II). Microfibrils of molecules displaced by

[6] M. Kurkinen, personal communication.

one-third of their thread length could be periodically stabilized by disulfide-linked segment junctions. Steric considerations preclude a tightly packed fiber of many molecules width, but the hexameric carboxyl (NCl) junctions could also be accommodated at the surface of a microfibril in which only a few molecules lie side-by-side.

We tested this concept numerically and found two arrangements with striking relationships. Consider a pair of antiparallel, homotrimeric *Drosophila* collagen molecules, [1] and [2], overlapping near their amino ends by their [A1-B] intervals, and disulfide-linked through them (Fig. 9*A*). Call this a dimer, [1,2]. Lay two dimers, [1,2] and [3,4], next to each other but mutually displaced by 489 residues. This positions interval [A2-C] of molecule [4] over interval [D2-F] in the parallel molecule [2] of the first dimer. Intermolecular disulfide bonds can therefore be made between cysteine residues [4,A2] and [2,D2], and between [4,C] and [2,F]. The next dimer, [5,6], is added in the same direction, again displaced by 489 residues. This aligns interval [A2-C] of molecule [6] with interval [D2-F] of molecule [4]. Further dimers can be added infinitely in both directions. This arrangement allows all 9 thread-cysteine residues of each chain to pair with other cysteines, within a maximum discrepancy of two residue positions, *i.e.* within the stagger of chains in a collagen helix. Fig. 9*B* illustrates how molecule [4] of Fig. 9*A* could form disulfide links with five adjacent molecules. Cysteines [A3], [D2], and [E] of three different molecules come into mutual alignment. Linkage could occur, as the three molecules represent the cysteines of nine chains at this location. Similarly, the extensive intermolecular disulfide bonding does not eliminate intramolecular disulfide links. Furthermore, additional disulfide links could still be made with other basement membrane components or with other microfibrils of *Drosophila* collagen.

Several other regular microfibrillar arrangements are possible, but in none of them are all the cysteine residues so well matched for interfibrillar disulfide bridging. Alignment of helix imperfections across one microfibril will make that structure more flexible. This arrangement is similar to the previous one, except that the paired, antiparallel [A1-B] junction is replaced by one of [A1-B-C] pairs (Fig. 10). Each of the borders of this segment junction approximately coincides with imperfections in all molecules lateral to it, and the boundaries of these imperfections either overlap or are at most one (Gly-*X*-*Y*) triplet apart.

In addition, each carboxyl end of a molecule in the above microfibrils can be linked to the carboxyl end of another molecule in the usual antiparallel arrangement formed by a vertebrate hexameric NCl junction. Several assumptions are implicit in these suggestions for supramolecular arrangements. First, that residue number and physical location are synonymous. Overall, the imperfections do not greatly change the thread length observed in the electron microscope from that expected for this number of residues in collagen helix. Furthermore, any effects would be the same in all molecules of a microfibril and are likely to average out. Second, it is assumed that the amino end of the molecule, the peptide up to cysteine [A1], does not interfere in the assembly. Independently of the length of this peptide after processing, it can be accommodated at the surface of the microfibril. The third assumption is that the carboxyl junctions do not interfere. They can also be accommodated at the surface of the micro-

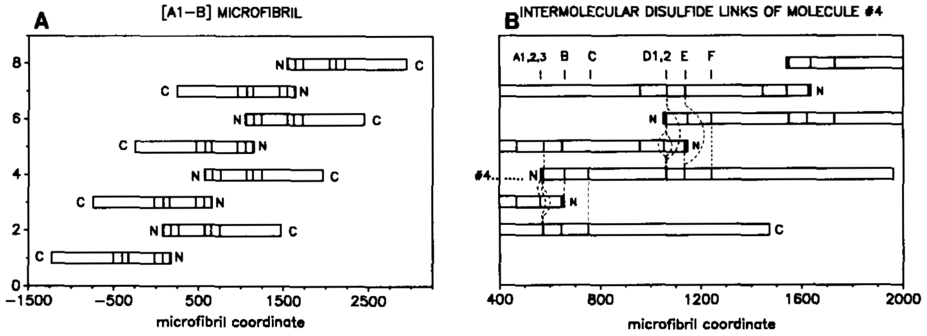

FIG. 9. *A*, arrangement of *Drosophila* procollagen IV molecules in microfibril based on antiparallel [A1-B] segment junctions. Each *horizontal bar* indicates the collagen thread portion of a three-chained molecule in the shown amino (*N*) to carboxyl (*C*) orientation. Except at the C end of the bar, each *vertical line* indicates the location of cysteine residues. Due to the scale used, the bars representing the A1, A2, and A3 residues cannot be distinguished, and those of D1 and D2 are also merged. Each pair of antiparallel molecules is displaced relative to adjacent pairs in the flat diagram, by 489 residues. At each carboxyl (*C*) end of the collagen thread another antiparallel molecule is joined, in line, through a hexameric NCl, carboxyl-carboxyl junction (not shown). See text for further explanation, and *B* for a diagram, at higher resolution, to illustrate the juxtapositioning of cysteine residues in different molecules. In this arrangement all 9 cysteine residues of the collagen thread portion of a chain can be disulfide linked to other chains. The microfibril coordinate scale is in amino acid residues, with the starting residue of molecule 2 placed at the zero position of the *abscissa* scale. For reference, the eight molecules are assigned the numbers shown on the *ordinate* scale. *B*, potential disulfide linkages of one *Drosophila* procollagen IV molecule to other molecules in a microfibril based on antiparallel [A1-B] segment junctions. The same arrangement of molecules as in *A* is shown at higher resolution, to diagram the potential disulfide links that can be made from cysteine residues of molecule 4. The positions of these cysteines of molecule 4 are indicated by the alphanumeric scale *A1* through *F*. The 9 cysteine residues of the thread portion of any molecule in the fibril are matched by cysteine residues in other molecules to within 2 residues. This is within the mutual stagger of ±1 residue in which a given residue in one of the strands of a homotrimeric collagen molecule is relative to the identical residues in the two companion α chains. Disulfide links are indicated by *dashed lines*, which are curved in the diagram to indicate linkage to molecules that are adjacent in the three-dimensional microfibril, but not in its flat representation. Note that one molecule, 4, can form disulfide links to five adjacent molecules. The same holds for the complementary, inverted set of molecules that would fit with that shown here by carboxyl-carboxyl junctions. The diagram illustrates the following disulfide links: {4,A1-3,B}; {4,A2-2,D1}; {4,A3-2,D2} and {4,A3-5,E}; {4,B-3,A1}; {4,C-2,F}; {4,D1-6,A2}; {4,D2-6,A3} and {4,D2-7,E}; {4,E-5,A3} and {4,D2-7,E}; {4,E-5,A3} and {4,E-7,D2}; {4,F-6,C}.

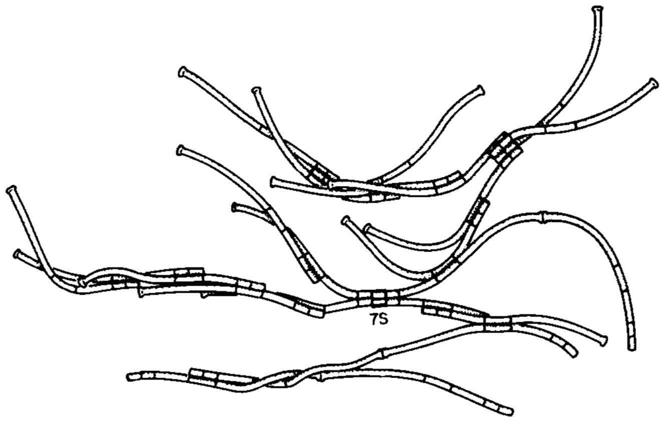

FIG. 11. **Diagram to illustrate the variety of possible arrangements of *Drosophila* procollagen IV molecules that may be stabilized by segment junctions.** Each three-stranded molecule is drawn as a flexible cylinder with a protrusion at one end to denote the carboxyl (NC1) domain. The other end of each cylinder is at the A1 cysteine residue. The length of each cylinder, without protrusion, denotes the length of the collagen thread portion of each molecule. *Ring marks* around each cylinder, from the non-knob end, denote the locations of cysteine residues B, C, D, and F. Cysteine residues A1, 2, and 3 are assumed to be at the amino end. Cysteine residue E is not indicated. Segment junctions between both parallel and antiparallel arrangement of intervals of the type [A-B], [A-C], and [D-F] are illustrated. A tetrameric junction of the vertebrate 7 S type is indicated, and to the left of it a portion of a microfibril. The complimentary antiparallel portion of this microfibril has been omitted for the sake of clarity.

fibril, and this should be facilitated by a prominent flex point in the thread near the carboxyl end, that is evident in electron micrographs (15, 50).

Networks—Drosophila basement membrane collagen IV can potentially participate in many kinds of molecular networks. We first define elements that are needed in a network. A minimal requirement is a set of continuous threads and some form of junctions. A molecule participates in a thread when it is joined to two other molecules at different places along its length. These places do not need to be at the ends of the molecules. A junction between two molecules is fixed by one or more covalent bonds. An *X* junction between two threads may be formed by a topological entanglement, or by one or more covalent bonds, or by both. A *Y* junction occurs when the end of one thread is joined to some portion of another thread that is not one of its ends. A thread can be one or more molecules thick (Fig. 11).

The carboxyl (NCl) end of a molecule can form one of the junctions needed for a molecule to participate in a thread. While the hexameric NCl junction joins two antiparallel molecules into a thread, it might also form attachments to other molecules. Such a thread of two collagen molecules is extended when any other portion of the molecules is covalently linked to another collagen molecule, for our purposes by disulfide and/or lysine-derived bonds. Segment junctions are other plausible links between collagen molecules, and good candidates for 30 and 60-nm long junctions are listed in Table III. The homotrimeric molecules provide multiple cysteine residues that can link several molecules at one junction. The additional molecules may be part of the same thread or of another thread. Further noncovalent and covalent interactions with other components of basement membranes will occur, and they may facilitate the formation of a collagen network, bias the choice of alternative networks, and both limit and stabilize the final product.

Wider Implications—Electron microscopy has given various indications of fine fibrillar elements in basement membranes, although overt fibers are missing. The collagen scaffold of basement membranes has been thought of as 1) an array of microfibrils (13, 32), 2) superimposed layers of a monomolecular thread network (11, 33), and 3) linked sheets of polygons made up from overlapping dimeric units (14, 49). From mammalian basement membranes the following two key junctional elements have been isolated: the 7 S segment junction of the amino ends of four collagen molecules and the NCl hexamer representing a pair of joined carboxyl ends. Kühn and associates (11, 13) proposed a general two-dimensional network of monomolecular threads linked by these junctions and interdigitated with adjacent network layers. Furthmayer and co-workers (49) extracted antiparallel, dimeric collagen IV molecules from basement membranes and studied their reassociation *in vitro*. They interpreted the resulting aggregates as sheets of polygons with varying degrees of statistical regularity and proposed that each polygon is made up of lateral associations of the antiparallel collagen IV dimers. Furthmayer and associates relegate the 7 S amino segment junctions to a relatively unimportant junctional role between layers of polygons, while Kühn *et al.* (11, 33) see them as key elements of their networks.

We propose a general formulation of basement membrane collagen networks which includes parts of the above models and some others. *Drosophila* procollagen IV clearly contains the elements of multiple segment junctions, which are not confined to the amino ends, as well as the evolutionarily conserved carboxyl junctions, and the potential for microfibril formation (Fig. 11). The existence of junctional elements at both ends of each molecule allows the formation of monomolecular threads.

General experience shows that an inherent property of collagen helices is their ability for lateral association, and this strengthens collagen as a tensile system. While segment junctions stabilize lateral molecular associations, imperfections of collagen helices destabilize them in at least two ways: by interfering in lateral fit and by providing flexibility for divergence of molecular threads. On the other hand, during multimolecular association, flexibility, combined with a helically preferred twist, is likely to lead to topological entanglements that would provide additional, pliable network junctions. The macroscopic equivalents of such junctions are fundamentals of knitted fabrics. Recent electron microscopy of basement membranes shows such thread entanglement (34). The conserved locations of half of the imperfections of helix between *Drosophila* and vertebrates points toward some feature of supramolecular assembly that depends on these imperfections and which remains to be understood.

A collagen microfibril stabilized by disulfide-bound segment junctions may represent a primitive but versatile primordial assembly that was subsequently refined into either the massive structures of vertebrate fibrous collagens, or into specialized basement membrane networks. Conceptually, such a cross-linked microfibril could be pulled asunder laterally into a micronetwork. Recent electron microscopic analysis of tendon indicates that collagen fibrils split and fuse (35). This is topologically related to a collapsed network. We conclude that there are structural continuities between fibers and networks.

In conclusion, we propose that basement membranes contain diverse associations of collagen IV molecules and that discordant views of microfibrils and partly fused monomolecular collagen threads are more a matter of classification than substance. We envisage associations that may appear locally as linked microfibrils and at an adjacent site more as monomolecular thread networks. It is likely that variations in the structure of basement membranes are strongly influenced by the noncollagenous components and will vary in different tissues. Perhaps key elements of basement membrane collagens are a combination of several changes in the direction of the axis of a flexible collagen thread, due to helix imperfections, and the ability to form several types of junctions between molecules. While *Drosophila* procollagen IV may be a particularly versatile form of collagen, vertebrate collagens sharing various aspects of these properties probably exist.

Acknowledgments—We thank Dr. M. Kurkinen and Dr. K. Kühn for providing us with sequence information prior to publication. J. H. F. thanks Dr. Norman Davidson and Dr. Karl Sirotkin for patiently helping to initiate this project.

REFERENCES

1. Blumberg, B., MacKrell, A. J., Olson, P. F., Kurkinen, M., Monson, J. M., Natzle, J. E., and Fessler, J. H. (1987) *J. Biol. Chem.* **262,** 5947–5950
2. Kurkinen, M., Bernard, M. P., Barlow, D. P., and Chow, L. T. (1985) *Nature* **317,** 177–179
3. Sakurai, Y., Sullivan, M., and Yamada, Y. (1986) *J. Biol. Chem.* **261,** 6654–6657
4. Soininen, R., Tikka, L., Chow, L., Pihlajaniemi, T., Kurkinen, M., Prockop, D. J., Boyd, C. D., and Tryggvason, K. (1986) *Proc. Natl. Acad. Sci. U. S. A.* **83,** 1568–1572
5. Trueb, B., Grobli, B., Spiess, M., Odermatt, B. F., and Winterhalter, K. H. (1982) *J. Biol. Chem.* **257,** 5239–5245
6. Haralson, M. A., Federspiel, S. J., Martinez-Hernandez, A., Rhodes R. K., and Miller, E. J. (1985) *Biochemistry* **24,** 5792–5797
7. Bächinger, H. P., Doege, K. J., Petschek, J. P., Fessler, L. I., and Fessler, J. H. (1982) *J. Biol. Chem.* **257,** 14590–14592
8. Hofmann, H., Voss, T., Kühn, K., and Engel, J. (1984) *J. Mol. Biol.* **172,** 325–343
9. Timpl, R., Oberbaumer, I., Von Der Mark, H., Bode, W., Wick, G., Weber, S., and Engel, J. (1985) *Ann. N. Y. Acad. Sci.* **460,** 58–72
10. Brazel, D., Oberbaumer, I., Dieringer, H., Babel, W., Glanville, R. W., Deutzman, R., and Kühn, K. (1987) *Eur. J. Biochem.* **168,** 529–536
11. Timpl, R., Wiedemann, H., Van Delden, V., Furthmayr, H., and Kühn, K. (1981) *Eur. J. Biochem.* **120,** 203–211
12. Weber, S., Engel, J., Wiedemann, H., Glanville, R. W., and Timpl, R. (1984) *Eur. J. Biochem.* **139,** 401–410
13. Kefalides, N. A., Howard, P., and Ohno, N. (1985) in *Basement Membranes* (Shibata, S., ed) pp. 73–88, Elsevier Scientific Publishing Co., Amsterdam
14. Yurchenko, P. D., and Furthmayr, H. (1984) *Biochemistry* **23,** 1839–1850
15. Fessler, J. H., Lunstrum, G., Duncan, K. G., Campbell, A. G., Sterne, R., Bächinger, H. P., and Fessler, L. I. (1984) in *The Role of Extracellular Matrix in Development* (Trelstad, R., ed) pp. 207–219, Alan R. Liss, New York
16. Fessler, L. I., Campbell, A. G., Duncan, K. G., and Fessler, J. H. (1987) *J. Cell Biol.* **105,** 2383–2391
17. Campbell, A. G., Fessler, L. I., Salo, T., and Fessler, J. H. (1987) *J. Biol. Chem.* **262,** 17605–17612
18. Campbell, A. G. (1986) Ph.D. thesis, University of California at Los Angeles.
19. Lunstrum, G. P. (1980) Ph.D. thesis, University of California at Los Angeles.
20. Monson, J. M., Natzle, J., Friedman, J., and McCarthy, B. J. (1982) *Proc. Natl. Acad. Sci. U. S. A.* **79,** 1761–1765
21. Ramachandran, G. N., and Ramakrishnan, C. (1976) in *Biochemistry of Collagen* (Ramachandran, G. N., and Reddi, A. H., ed) pp. 45–84, Plenum Publishing Co., New York
22. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K., and Efstratiadis, A. (1978) *Cell* **15,** 687–701
23. Boedtker, H., Fuller, F., and Tate, V. (1983) *Int. Rev. Connect. Tissue Res.* **10,** 1–63
24. Yamada, Y., Avvedimento, V. E., Mudryj, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I., and de Crombrugghe, B. (1980) *Cell* **22,** 887–892
25. Dedhar, S., Ruoslahti, E., and Pierschbacher, M. D. (1987) *J. Cell Biol.* **104,** 585–593
26. Ruoslahti, E., and Pierschbacher, M. D. (1986) *Cell* **44,** 517–518
27. Dickerson, R. E. (1971) *J. Mol. Evol.* **1,** 26–45
28. Soininen, R., Chow, L., Kurkinen, M., Tryggvason, K., and Prockop, D. J. (1986) *EMBO J.* **5,** 2821–2823
29. Burgeson, R. E., Morris, N. P., Murray, L. W., Duncan, K. G., Keene, D. R., and Sakai, L. Y. (1985) *Ann. N. Y. Acad. Sci.* **460,** 47–57
30. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132
31. Duncan, K. G. (1985) Ph.D. thesis, University of California at Los Angeles
32. Veis, A., and Schwartz, D. (1981) *Collagen Rel. Res.* **1,** 269–286
33. Kühn, K., Glanville, R. W., Babel, W., Quan, R. Q., Dieringer, H., Voss, T., Siebold, B., Oberbaumer, I., Schwarz, U., and Yamada, Y. (1985) *Ann. N. Y. Acad. Sci.* **460,** 14–24
34. Yurchenco, P. D., and Ruben, G. C. (1987) *J. Cell Biol.* **105,** 2559–2568
35. Birk, D. E., and Trelstad, R. L. (1987) *Proc. Electronmicroscopy Soc. Am.* **45,** 574–577
36. Young, R. A., and Davis, R. W. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80,** 1194–1198
37. Huynh, T. V., Young, R. A., and Davis, R. W. (1985) in *DNA Cloning: a Practical Approach* (Glover, D., ed) Vol. I, pp. 49–78, IRL Press Ltd., Oxford
38. Leder, P., Tiemeir, D., and Enquist, L. (1977) *Science* **196,** 175–177
39. Yanisch-Perron, C., Vierira, J., and Messing, J. (1985) *Gene (Amst.)* **33,** 103–119
40. Tabor, S., and Richardson, C. C. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84,** 4767–4771
41. Staden, R. (1980) *Nucleic Acids Res.* **8,** 3673–3694
42. Devereaux, J., Haeberli, P., and Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395
43. Dayhoff, M. O. (1976) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed) Vol. 5, Supplement 2, pp. 4–6, National Biomedical Research Foundation, Washington, D. C.
44. Dayhoff, M. O., Barker, W. C., and Hunt, L. T. (1983) *Methods Enzymol.* **91,** 524–545
45. Kanehisa, M. I. (1982) *Nucleic Acids Res.* **10,** 183–196
46. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517
47. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472
48. Keller, E. B., and Noon, W. A. (1985) *Nucleic Acids Res.* **13,** 4971–4981
49. Yurchenko, P. D., Tsilibary, E. D., Charonis, A. S., and Furthmayr, H. (1986) *J. Histochem. Cytochem.* **34,** 93–102
50. Lunstrum, G. P., Bächinger, H. P., Fessler, L. I., Duncan, K. G., Nelson, R. E., and Fessler, J. H. (1988) *J. Biol. Chem.* **263,** 18318–18327

Miniprint supplement to Blumberg, MacKrell and Fessler
Drosophila Basement Membrane Procollagen α1(IV):
Complete cDNA Sequence, Genomic Structure and General Implications
for Supramolecular Assemblies

MATERIALS AND METHODS

Construction and screening of cDNA libraries

cDNA was prepared from a *Drosophila melanogaster* Kc cell line known to produce basement membrane components (15) as described (1). The cDNA was size selected and ligated into the vectors λgt11 (36), λgt10 (37) and λgtWES (38). These libraries contain respectively 10^6, 2.5×10^5 and 6×10^3 clones. We prepared [^{32}P]-labeled RNA (1) from the cloned *Drosophila* genomic DNA fragment DCG1A (20) and screened the λgt10 library at high stringency (1). Clones giving duplicating positive signals were plaque purified and subcloned into the plasmid vector Bluescribe (Stratagene Cloning Systems, San Diego, CA) for restriction-mapping. Clones λCDA3, λCDA10, λCDA11, λCDB5, and λCDB8 contained the largest inserts and were chosen for further study, the plasmid clones containing the inserts from these clones are called, respectively, pBB103, pAM101, pAM102, pBB122, and pAM106.

DNA sequence analysis

Appropriate restriction fragments were subcloned into the vectors M13mp18, M13mp19 (39) or Phagescript (Stratagene). The sequencing strategy that was used is shown in Fig 1. The cDNA was entirely sequenced on both strands by the dideoxy method using α-[^{35}S]-labeled dATP (New England Nuclear) and avian myeloblastosis virus reverse transcriptase (Seikagaku America, St. Petersburg, FL) as described (1). Portions of the A-T rich intron-containing sequences were sequenced using the Sequenase system (United States Biochemical Corp., Cleveland, OH) according to the protocol supplied by the manufacturer. Briefly, this system is based on the dideoxy method but utilizes a modified bacteriophage T7 DNA polymerase (40) which is relatively insensitive to template sequence and secondary structure. The DNA sequence was maintained and aligned during the sequencing using Staden's DB-system (41). DNA and amino acid sequences were analyzed using various programs of the University of Wisconsin Genetics Computer Group (42), the National Biomedical Research Foundation (43, 44), and the Los Alamos Portable sequence homology package (45).

Isolation of genomic clones

To isolate genomic DNA clones for *Drosophila* proα1(IV) collagen, *in vitro* [^{32}P]-labeled runoff RNA probes were prepared from the 5′ (pAM102) and 3′ (pBB122) ends of the cDNA and used to screen a *Drosophila melanogaster* genomic DNA library (22) at high stringency (1). Positive clones were plaque purified and restriction-mapped with EcoRI. Southern blot analysis (46) of the EcoRI digested phage DNA confirmed the ordering of the restriction fragments and was used to correlate the cDNA map to the genomic map.

RESULTS

Intron/exon boundaries

The sequences at the 5′ and 3′ intron/exon boundaries agree with the eukaryotic consensus sequences (47) and are shown in Figure 5. Furthermore, the consensus 3′ splice signal proposed for *Drosophila*, C/T-T-A/G-A-C/T (48), is present 16 - 40 bp 5′ to each 3′ splice acceptor site in the introns. The sequences are shown in Fig 5.

All but one of the exons sequenced from the triple-helical domains of mouse α1(IV) (3) and mouse α2(IV) (2)[5] begin with a 2/3 intact glycine codon and correspondingly end with a 1/3 intact glycine codon. This is distinct from the exons of interstitial collagens which always begin with an intact glycine codon (23). In contrast, of the 4 exons in the triple-helical domain of *Drosophila* proα1(IV), two begin with 2/3 intact glycine codons, one with a 1/3 intact glycine codon, and one with an intact glycine codon. All four glycine codons are GGT.

TABLE 1

IMPERFECTIONS OF THE GLY-X-Y SEQUENCES IN THE COLLAGEN THREAD OF *DROSOPHILA* PROα1(IV)

#	Sequence	Length	NH2-dist	NC1-dist
1	KNCTAGYAGCVPKCIAEK	18	71	1475
2	AKEN	4	203	1343
3	AK	2	210	1336
4	CY	2	263	1283
5	KP	2	268	1278
6	ASSFPVKPTHTVM	13	282	1264
7	R	1	397	1149
8	DGA	3	437	1109
9	GY	2	491	1055
10	KL	2	529	1017
11	CSSCRA	6	570	976
12	ALCDLSLIEPLK	12	642	904
13	IK	2	720	826
14	CALDEIKMPAK	11	751	795
15	RP	2	852	694
16	SEK	3	1061	485
17	VH	2	1097	449
18	ATVPDIR	7	1229	317
19	IK	2	1295	251
20	ESRLV	5	1330	216
21	PPP	3	1461	85

The position of the amino end of each imperfection is listed as the residue number from the start Met and from the collagen thread/NC1 junction. The number of residues in each imperfection is also given. See text for definition of imperfection.

TABLE 2

INTERVALS BETWEEN CYSTEINE RESIDUES OF *DROSOPHILA* PROCOLLAGEN IV

	Interval	Residues		Difference	
		from amino	from thread/NC1	aa	nm
(1)	[A1-B]	73, 168	1471, 1376	95	28
	[A2-B]	80, 168	1464, 1376	88	26
	[A3-B]	84, 168	1460, 1376	84	24
	[B-C]	168, 263	1376, 1281	95	28
(2)	[A1-C]	73, 263	1471, 1281	190	55
(3)	[A2-C]	80, 263	1464, 1281	183	53
	[D1-F]	570, 751	974, 793	181	53
	[A3-C]	84, 263	1460, 1281	179	52
	[D2-F]	573, 751	971, 793	178	52
(4)	[A2-D1]	80, 570	1464, 974	490	
	[A3-D2]	84, 573	1460, 971	489	
	[C-F]	263, 751	1281, 793	488	
			average:	489	142
	[B-E]	168, 644	1376, 900	476	138

The cysteine residues are labeled alphanumerically from the amino end of the procollagen chain: A1, A2, A3, B, C, D1, D2, E and F. The residue positions are separately listed from the start Met residue, and from the junction of the end of the collagen-thread (Gly-X-Y) sequence and the beginning of the NC1 region. The length of each interval between cysteine residues is given in number of amino acids and in physical length calculated for an assumed 0.29 nm translation per residue.

TABLE 3

POTENTIAL SEGMENT JUNCTIONS OF *DROSOPHILA* PROCOLLAGEN IV

```
1.     [A2---K------K---B]
        80  83      165 168

2.     [A3---K---K---K---K---B]
        84  88  118 133 165 168

3.   K---[A1------K---B]
    71    73      165 168

4.      K---[B------K------K------C]---K
       165  168     196    235    263  268

5.     K--[A1---K---K---K---K----B----K---K---K---K---C]--K
       48  73   88  103 121 133  168  204 214 235 247  263 288

6.      K----[D2---K----K------K----K---F]----K
       557   573  578  592    733  745 751   757
```

Intervals between alphanumerically labeled cysteine residues are shown that are likely candidates as components of segment junctions. Lysine residues that are potential bridgeheads for lysine-derived cross-links are indicated as K, and the residue positions are given from the amino end. Except for interval (3), all the shown intervals have an axis of symmetry with respect to potential cross-linking bridgeheads, i.e. an antiparallel pair of a given interval may form a multiply stabilized segment junction.

```
INTRON 5' DONOR    3' SPLICE SIGNAL      3' ACCEPTOR

1 CCT/GTAAGG---293 BP---TTAAA---10 BP---TCTCTGGCTCATCTTTCAG/GCC
2 GCG/GTAGAA---449 BP---CTAAC---10 BP---TTCGTCGACTTCTTTGCAG/GCT
3 GCT/GTAAGA--- 28 BP---TTGAA--- 4 BP---ATACTCCTTTTCCCCGCAG/CAA
4 CGC/GTAAGT--- 34 BP---TTGAA--- 9 BP---ATGAAATATATACGTACAG/GGT
5 TGG/GTAAGT--- 87 BP---TTAAT--- 6 BP---TAATTTTCGTTCGTCCTAG/TTA
6 CCG/GTAAGT--- 22 BP---TTGAT--- 8 BP---TTTGGTTTCATTTGTTCAG/GTG
7 AAG/GTGAGT--- 30 BP---TTCAT---20 BP---CGTTATTTAACCCATTCAG/GTC
8 ATG/GTGAGG---230 BP---CTAAT--- 7 BP---CAACCATTTCCTATAATAG/CAC
```

Figure 5. Intron/exon splice junctions in the *Drosophila* proα1(IV) gene. The 5′ donor and 3′ acceptor sequences for the *Drosophila* proα1(IV) are shown flanking the putative 3′ splice signals. These sequences correspond closely with the splice donor and acceptor consensus sequences (47) and the splice signal consensus (48).

```
A - COMPARISON OF DROSOPHILA AND HUMAN ALPHA1(IV) 7S REGIONS

       76        86        96        106       116       126       136
DROS   CTA GYAGCVPKCI AEKGNRGLPGPLGPTGLKGENGFPGMEGPSGDKGQKGDPGPYGDRGDKGERGSPG
       |  |  |       || |||||  |  |  |  |  |  || | ||  | || | |  |  | |
HUMA1  CAGSGCGKCDCHGVKGQKGERGLPGLQGVIGFPGMQGPEGPQGPPGQKGDTGEPGLPGTKGTRGPPGASG
       34        44        54        64        74        84        94

       144       154       164       174       184       194       204
DROS   LHGQAGVPGVQGPAGNPGAPGINGKDGCDGQDGIPGLEGLSGMPGPRGYAGQLGSKGEKGEP AKENGDY
         |  | ||  |  | || ||| |  |  |  |  |  || |  |  |  |  | ||  ||      |
HUMA1  YPGNPGLPGIPGQDGPPGPPGIPGCNGTKGERGPLGPPGLPGFAGNPGPPGLPGMKGDPGEILGHVPGML
       104       114       124       134       144       154       164

       213       223       233       243       253       263       273
DROS   AKGEKGEPGWRGTAGLAGPQGFPGEKGERGDSGPYGAKGPRGEHGLKGEKGASCYGPM KPGAPGIKGEK
        ||| | ||  || |  |  |  |  |  |  || |  || |  | ||  | |  ||    |  |  |
HUMA1  LKGERGFPGIPGTPGPPGLPGLQGPVGPPGFTGPPGPPGPPGPPGEKGQMGLSFQGPKGDKGDQGVSGPP
       174       184       194       204       214       224       234

B - COMPARISON OF DROSOPHILA AND MOUSE ALPHA2(IV) 7S REGIONS

       63        73        83        93        103       113       123
DROS   SAGVARGDLPPKNCTAGYAGCVPKCIAEKGNRGLPGPLGPTGLKG  EM GFPGMEGPSGDKGQKGDPGP
         ||  | |          ||   |  ||| || ||  || |  |       ||||  |   ||||  | |||
MOUA2  LGGVKKLDVPCGGRDCS GGC  QCYPEKGARGQPGAVGPQGYNGPPGLQGFPGLQGRKGDKGERGVPGP
       14        24        34        44        54        64        74

       130       140       150       160       170       180       190
DROS   YGQRGDKGERGSPGLHGQAGVPGVQGPAGNPGAPGINGKDGCDGQDGIPGLEGLSGMPGPRGYAGQLGSK
        |  || | ||  |  |  | ||  |  |  | ||  |  |  |  |  |  |  |  | ||  |  | | |
MOUA2  TGPKGDVGARGVSGFPGADGIPGHPGQGGPRGRPGYDGCNGTRGDAGPQGPSGSGGFPGLPGPKDQ GQK
       81        91        101       111       121       131       141

       200       210       220       230       240       250       260
DROS   GEKGEPAKENGD YAKGEKGEPGWRGTAGLAGPQGFPGEKGERGDSGPYGAKGPRGEHGLKGEKGASCYG
       ||      ||  | |  || ||||  |  |  |  |  |  |  |  |  |  |  || |  |  |  | ||
MOUA2  GEPYALSKEDRDKY RGEPGEPGLVGYQGPPGRPGPIGQMGPMGAPGRPGPPGPPGPKGQPGNRGLGFYG
       150       160       170       180       190       200       210

       269       279       289
DROS   PM KPGAPGIKGEKGEPASSFPVKPTHTVMGP
            |  |  |  | |     | ||    |
MOUA2  QKGEKGDIGQPGPNGIPSDITLVGPTTSTIMP
       219       229       239
```

Figure 8. Comparison of the [A1-B-C] region of *Drosophila* proα1(IV) with (a) human α1(IV) and (b) mouse α2(IV) collagen chains for sequence homology. The sequences were aligned pairwise with the program SEQHP (45). In the alignment between *Drosophila* and human α1(IV) chains, 86/205 residues are identical (61 of these are glycines), and an additional 40/205 residues are conservatively substituted. In the alignment between *Drosophila* and mouse α2(IV) chains, 98/236 residues are identical (60 of these are glycines), and an additional 48 residues are conservatively substituted.

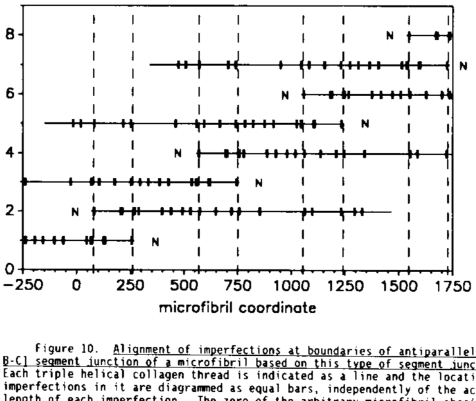

Figure 10. Alignment of imperfections at boundaries of antiparallel [A1-B-C] segment junction of a microfibril based on this type of segment junction. Each triple helical collagen thread is indicated as a line and the locations of imperfections in it are diagrammed as equal bars, independently of the actual length of each imperfection. The zero of the arbitrary microfibril abscissae coordinate is at the starting residue of molecule #2, and the length unit is an amino acid residue. The amino end of each molecule is indicated as N. Dashed lines give positions of laterally aligned imperfections. The ordinate gives the reference number of each molecule.