Candidate Cell and Matrix Interaction Domains on the Collagen Fibril, the Predominant Protein of Vertebrates*§

Shawn M. Sweeney, Joseph P. Orgel, Andrzej Fertala, Jon D. McAluliffe, Kevin R. Turner, Gloria A. Di Lullo, Steven Chen, Olga Antipova, Shiamalee Perumal, Leena Ala-Kokko, Antonella Forlino, Wayne A. Cabral, Aileen M. Barnes, Joan C. Marini, and James D. San Antonio

From the Cardiovascular Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, the Center for Synchrotron Radiation Research and Instrumentation, Department of Biological, Chemical, and Physical Sciences, Illinois Institute of Technology, Chicago, Illinois 60616, the Department of Dermatology and Cutaneous Biology, Thomas Jefferson University, Philadelphia, Pennsylvania 19107, the Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia 19104, Pennsylvania, the Chicago Medical School, North Chicago, Illinois 60064, the Collagen Research Unit, Biocenter and Department of Medical Biochemistry and Molecular Biology, University of Oulu, Oulu, Finland, the Connective Tissue Gene Tests, Allentown, Pennsylvania 18103, the Department of Biochemistry A. Castellani, University of Pavia, Pavia, Italy, the Bone and Extracellular Matrix Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, and the Cardeza Foundation for Hematologic Research and Department of Medicine, Thomas Jefferson University, Philadelphia, Pennsylvania 19107

Type I collagen, the predominant protein of vertebrates, polymerizes with types III and V collagens and non-collagenous molecules into large cable-like fibrils, yet how the fibril interacts with cells and other binding partners remains poorly understood. To help reveal insights into the collagen structure-function relationship, a data base was assembled including hundreds of type I collagen ligand binding sites and mutations on a two-dimensional model of the fibril. Visual examination of the distribution of functional sites, and statistical analysis of mutation distributions on the fibril suggest it is organized into two domains. The "cell interaction domain" is proposed to regulate dynamic aspects of collagen biology, including integrin-mediated cell interactions and fibril remodeling. The "matrix interaction domain" may assume a structural role, mediating collagen cross-linking, proteoglycan interactions, and tissue mineralization. Molecular modeling was used to superimpose the positions of functional sites and mutations from the two-dimensional fibril map onto a three-dimensional x-ray diffraction structure of the collagen microfibril in situ, indicating the existence of domains in the native fibril. Sequence searches revealed that major fibril domain elements are conserved in type I collagens through evolution and in the type II/XI collagen fibril predominant in cartilage. Moreover, the fibril domain model provides potential insights into the genotype-phenotype relationship for several classes of human connective tissue diseases, mechanisms of integrin clustering by fibrils, the polarity of fibril assembly, heterotypic fibril function, and connective tissue pathology in diabetes and aging.

Type I collagen is the most abundant protein in humans and other vertebrates, comprising much of the fibrous extracellular matrix scaffold of bones, tendons, skin, and many other tissues (1–4). In general, type I collagen and its binding partners are proposed to provide mechanical strength and form to tissues. Collagenous scaffolds are laid down and remodeled by cells and are also a predominant substrate for cell interactions, migration, and differentiation. Consequently, various debilitating human diseases are associated with type I collagen mutations, including osteogenesis imperfecta (OI, brittle bone disease), Ehlers Danlos syndrome, vascular disorders, and others (3, 5). Type I collagen is also employed in human medicine as hemostatic sponges and implants to repair wounds and in tissue engineering applications as scaffolds (6).

Type I collagen is synthesized in the endoplasmic reticulum as α1 and α2 procollagen chains, each encoded by separate genes that are translated into proteins somewhat longer than 1000 amino acid residues (3, 7). Nucleation domains on the C-terminal propeptide promote the polymerization of two α1 and one α2 chains into the procollagen triple helical monomer (Fig. 1, A and B). The triple helical domain of procollagen is composed of contiguous glycine-X-Y tripeptide repeats, with the obligate glycine in the first position as its side chain is the only one small enough to fit within the coiled-coil of the triple helix. Extracellularly, N- and C-proteinases remove the globular termini of procollagen, and every ~67 nm along the fiber axis, five monomers assemble in a quarter-staggered fashion to form part of the supramolecular “helix,” the microfibril. Each microfibril, the proposed subunit of the collagen fibril (Fig. 1, C–E), and its immediate microfibrillar neighbors are connected by N- and C-terminal intermolecular cross-links. The basic repeating

* This work was supported, in whole or in part, by National Institutes of Health Grants AR048544 (to A. F.), AHA 0435339Z, NIH RR08630, and NSF 0644015 (to J. P. R. O. O.), and NIH AR049604 and HL053590 (to J. S. A.). 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.

§ The on-line version of this article (available at http://www.jbc.org) contains supplemental text, equations, and Fig. S1.

1 To whom correspondence should be addressed: Thomas Jefferson University, Cardeza Foundation for Hematologic Research and Dept. of Medicine, Rm. 809 Curtis, 1015 Walnut St., Philadelphia, PA 19107-5099. Tel.: 215-955-6121; Fax: 215-923-3836; E-mail: james.sanantonio@jefferson.edu.

2 The abbreviations used are: OI, osteogenesis imperfecta; PG, proteoglycan; THP, triple helical peptide; MLBR, major ligand binding region; MMP, matrix metalloproteinase; SPARC, secreted protein, acidic, and rich in cysteine; FACIT, fibril-associated collagens with interrupted triple helices.
morphological structure of the fibril is the D-period, ~67 nm long, and composed of one overlap and one gap zone. Each D-period contains the complete monomer sequence derived from overlapping consecutive elements of five monomers (Fig. 1C, box). Other collagens, proteoglycans (PGs), and matrix macromolecules may further assemble with the fibril to impart tissue-specific properties to the heterotypic polymer (2, 4, 8). However, it is unclear how cells, matrix molecules, and other factors interact with the heterotypic collagen fibril. Therefore, a comprehensive structure-function model of the protein has been lacking.

To help elucidate type I collagen structure-function relationships, a data base of functional domains, ligand binding sites, and human disease-associated mutations mapping to type I collagen was previously assembled and presented as a two-dimensional map of the collagen fibril (9). It was reported that ligand-binding hot spots appear on collagen, and the positions of certain classes of mutations may correlate with disease phenotype. Subsequently the molecular, microfibrillar, and fibrillar structures of type I collagen have been determined through x-ray diffraction (10).

This study is a theoretical analysis of the interrelationships between hundreds of new functional domains, ligand-binding sites, and mutations on the collagen map. Moreover, collagen fibrils may be heterotypic, i.e. contain other collagens such as types III and V (15), whose arrangements in the fibril are incompletely understood, and that could impact fibril-ligand binding. However, despite the degree of uncertainty implicit in the map, and that some discrepancies exist in the data base (9), the approach followed here is to identify fundamental features of collagen structure and function based on numerous lines of evidence and by focusing on the best characterized functional elements of collagen.

**EXPERIMENTAL PROCEDURES**

**Ligand Binding Sites and Functional Domains**—Positions of binding sites and functional domains were obtained from the literature and are indicated by *labeled boxes* placed next to the relevant sequences. Primary literature references for sites indicated on an earlier version of the map appear in the supplemental materials. Recently reported sites are referenced in legend to Fig. 2. Zones of interactions between ligands and broad regions of collagen fibrils, as observed by electron microscopy, are indicated by *colored overlays*. Many binding site locations were approximated based on low resolution approaches, such as electron microscopy of macromolecular complexes. On the other hand, the use of collagen model triple helical peptides (THPs) (12, 13), in some cases, including novel peptide Toolkits composed of overlapping, 27-residue segments of collagen sequences (14), has allowed high resolution mapping of the positions of various ligand binding sites and functional domains on the collagen monomer.

*In vivo*, type I collagen-ligand interactions may depend upon the tissue source. Moreover, collagen fibrils may be heterotypic, i.e. contain other collagens such as types III and V (15), whose arrangements in the fibril are incompletely understood, and that could impact fibril-ligand binding. However, despite the degree of uncertainty implicit in the map, and that some discrepancies exist in the data base (9), the approach followed here is to identify fundamental features of collagen structure and function based on numerous lines of evidence and by focusing on the best characterized functional elements of collagen.

**Identifying Interrelationships between Sites**—Relevant relationships between binding sites and functional domains were identified in four ways. First, as sequences on the collagen monomer to which more than one ligand has been shown to bind, or that are near neighbors; second, as sequences that fall within the borders of a fibril region shown to bind a particular ligand; and third, as neighboring binding sites on adjacent monomers within the D-periodic packing scheme. In this last case it is proposed that interactions between ligands on adjacent monomers may be possible if their binding sites align vertically within the D-period, and if their two-dimensional structure allows them to simultaneously bind more than one triple helix and reach another ligand or ligand-binding site on neighboring monomer(s). Here, sites on two or more monomers are considered as aligned vertically if they overlap with or fall close to any interactions and as the predominant structural elements of vertebrate tissues.
axis drawn perpendicular to the long axes of the monomers, and joining monomers 1 and 5. Fourth, select data from the two-dimensional packing map were correlated with the known three-dimensional packing structure of collagen within the fibrillar context (10, 16).

Correlating Two-dimensional Collagen Map Data with a Three-dimensional Collagen Microfibril Model—The collagen microfibril model used in this study was composed from the packing structure of rat tendon type I collagen molecules observed in situ as described (10). Fiber diffraction data collected from native and derivatized (heavy atom labeled) rat tail tendons was used to solve the in situ structure of the collagen molecules and microfibril using multiple isomorphous replacement (16). A molecular model was constructed based on the primary sequences of the α1 and α2 chains of rat type I collagen, and the superhelical parameters were determined from collagen-like peptide crystallographic structure determinations. The electron density map representing the experimentally determined microfibril has a resolution of 0.516 nm in the direction of the fiber (collagen molecule) axis and 1.11 nm in the direction perpendicular to this. To determine the position of functional sites from the two-dimensional collagen map on the three-dimensional microfibril model, solvent-accessible surface calculation and rendering was performed using SPOCK (17) with the default probe size of 0.14 nm. Distances between surface sites were determined by line-of-sight measurement of atoms central to each site and located at the microfibril surface, assembled by using the instructions and coordinates provided by RCSB structure file 1YOF (10). To construct images from these analyses, the Ca “worm” traces of relevant portions of individual triple helices were marked using a semi-transparent surface rendering. The semi-transparent surface was then rendered in the appropriate colors to represent the positions of relevant functional sequences and binding sites. That rat and human collagen protein sequences are highly homologous justifies the approach of localizing functional domains of human type I collagen on the rat type I collagen microfibril.

Three-dimensional Modeling of Integrin-Collagen Interactions—Molecular modeling of α2β1 integrin I-domain binding to the collagen triple helix was performed on a Silicon Graphics (Octane) computer system using the SYBYL software package, version 7.2 (Tripos Inc.). A model of the collagen I triple helix was created (18) with a 36-amino acid peptide spanning the collagen primary sequences and the ClustalW function of MacVector 9.0 (Accelrys). In homology determinations between two proteins the number of homologous residues divided by the number of residues in the larger protein yielded the percent homology value. The find feature of the program was used to search for functional domain sequences of interest. Accession numbers for collagen sequences examined in this study include: Homo sapiens COL1A1, NM_000088; H. sapiens COL1A2, NM_000089; H. sapiens COL2A1, BC116449; H. sapiens COL3A1, NM_000090; H. sapiens COL5A1, NM_000093; H. sapiens COL5A2, NM_000393; H. sapiens COL5A3, NM_015719; H. sapiens COL11A1, J04177, J05407, NM_080630, NM_08062, NM_001854; H. sapiens COL11A2, NM_080681, NM_080680, NM_080679; Bos taurus COL1A1, BC105184; B. taurus COL1A2, NM_174520; Canis familiaris COL1A1, NM_01003090; C. familiaris COL1A2, NM_01003187; Xenopus laevis COL1A1, BC049829; X. laevis COL1A2, BC049287; Mus musculus COL1A1, NM_007742; M. musculus COL1A2, NM_007743; Rattus norvegicus COL1A1, XM_001081230; R. norvegicus COL1A2, NM_053356; Danio rerio COL1A1, NM_199214, and D. rerio COL1A2, NM_182968.

Statistics—Statistical analyses of mutation distributions on the collagen fibril appear under “Results” and in the supplemental materials.

RESULTS AND DISCUSSION

The collagen map (Fig. 2) consists of the protein sequences of the human α1 and α2 chains of type I collagen, arranged as they are proposed to occur in the fibril, upon which positions of structural landmarks, ligand binding sites, and disease-associated mutations are superimposed.

Ligand Binding Site Distribution—When the distribution of ligand-binding sites was examined on the map, more were found on the C-terminal half of the collagen monomer (Fig. 2), as noted previously (9). Moreover, three concentrations or “hot spots” of ligand binding are again evident, designated as major ligand binding regions (MLBRs) (dashed boxes 1–3, Fig. 2). The potential functional implications of some of the overlapping ligand binding sites were discussed before (9). A previously unobserved and less conspicuous hot spot appears for the first time on the revised map on and around sequence glycine-phenylalanine-hydroxyproline-glycine-glutamic acid-arginine (GFPGER)502–507 (Fig. 2, orange box), previously discovered as the predominant α1β1/α2β1/α11β1 integrin and cell binding
FIGURE 2. **Domain Model of the Collagen Fibril**. Protein sequences of the triple helix are shown (GenBankTM, α1(I) accession #NP000079.2 and α2(I) NP_000080.2; proline and hydroxyproline are designated as P, the latter being the third position in glycine-X-Y sequences. Note that Ala-459, α2(I) results from a single nucleotide polymorphism observed in some patient samples in the non-redundant protein database, whereas α2(I) Pro-459 appears in the NP_000080.2 sequence. Yet, the association of Ala to Pro-459 substitution mutations with intracranial aneurisms (26) justifies its
**Domain Model of the Collagen Fibril**

Inclusion here. Ligand binding regions are indicated by **rectangle boxes** adjacent to relevant sequences; ligand binding to type I procollagen or tropocollagen (gray), or to either α1(I) or α2(I) chains (unshaded). Ligand binding hot spots 1, 2, and 3 are enclosed by **dashed-lined black boxes**. **Highlighted sequences** include the α1(I)/α2(I) integrin binding site GFPGER**302–307** (maroon rectangle); MMP-1 cleavage site (green arrow); N- and C-terminal intermolecular cross-links (purple rectangles). Studies responsible for mapping many of the functional domains on collagen are cited in the first map publication (9) and in the supplemental materials. New sites on the revised map include endo180 (57), HSPG4 (58), micro-unfolding and N-anchor domains (59, 60), MMP-interaction domain (11), prolyl-3-hydroxylase substrate P986 (61), SPARC (62), phosphophoryn (63, 64), von Willebrand Factor (65), discoidin domain receptor 2 (66), and activities of THPs on angiogenesis (13), endothelial cell activation (67), and osteoblast differentiation (68). Mutations were from sources cited under the “Experimental Procedures.” This figure was modified from a previously published illustration (9).

| α1(I) integrin | α1(I) integrin-binding sites |
|----------------|-----------------------------|
| α2(I) integrin | α2(I) integrin-binding sites |
| α1(I)/β1 integrin | α1(I)/α2(I) integrin-binding sites |
| AD | Mutation associated with arterial dilatation |
| ANE | Mutation associated with aneurysm |
| CAD | Mutation associated with cerebral artery dilatation |
| CAF | Mutation associated with Caffey Disease (infantile cortical hyperostosis) |
| Cxv X-Hnk | Collagen V cross-link site |
| COMP | Cartilage oligomeric matrix protein binding sites |
| DPC | C-Proteinase cleavage site |
| CTW | Mutation associated with connective tissue weakness |
| DC integrin | Integrin binding site on denatured collagen |
| DDR2 | Proposed site of decorin receptor 2 binding |
| Decorin core | Decorin core protein binding site |
| DI | Mutations associated with Dentigeneisis imperfecta |
| [PS] HS PG | Proposed site of dermatan sulfate proteoglycan binding |
| EDS | Mutations associated with Ehlers-Danlos Syndrome |
| ENDL | Mutations associated with Ehlers-Danlos VII-like Syndrome |
| ENDO180 | Collagen binding factor |
| Fibri (+) | C-telopeptide fibrillogenesis nucleation domain |
| Fibri (-) | Fibrillogenesis inhibition site |
| G→X | Substitution mutation |
| GE-decorin | Guanidine-extracted decorin binding sites |
| Glycation | Glycation sites |
| HEP | Heparin: heparan sulfate proteoglycan binding site |
| HSPG4 | Potential heat shock protein 41 (thapsigargin) binding sites |
| HCA | Mutation associated with intracellular aneurysms |
| [PS] KS PG | Proposed sites of keratan sulfate proteoglycan binding |
| MFS | Mutation associated with Marfan Syndrome |
| MMP-1 | Matrix Metalloproteinase-1 cleavage site |
| MMP-1, 2, 13 | 1, 2, 13 |
| MMP Interaction Domain | Matrix Metalloproteinase Interaction Domain |
| N-ANCHOR DOMAIN | Proposed N-anchor Domain |
| NPC | N-Proteinase cleavage sites |
| OI | Mutations associated with Osteogenesis imperfecta |
| types 1, 2, 3, 4; 7H US denotes difficult diagnosis/unusual phenotype |
| OPA | Mutation associated with Osteopenia; 7 denotes difficult diagnosis/unusual phenotype |
| OPG | Mutations associated with Osteoporosis; 7 denotes difficult diagnosis/unusual phenotype |
| P-3-H | Site of Prolyl-3-hydroxylation |
| PHOSPH | Phosphorylating binding site |
| PHOSPHOPHORYL | Phosphorylating binding site |
| SPARC | Secreted protein acidic and rich in cystein binding site |
| TNP1/2, 11-β1 integrin | Triple helical peptide: binds integrin receptors |
| TnPiA/G | Triple helical peptide: inhibits angiogenesis |
| TnPiECA+/ | Triple helical peptide: promotes endothelial cell activation |
| TnPiFA+ | Triple helical peptide: promotes fibroblast adhesion, substrate for MMPs |
| TnPiH/AN- | Triple helical peptide: binds heparan, inhibits angiogenesis |
| ThPnOB-D | Triple helical peptide: supports osteoblastic differentiation |
| μ Unfolding | Proposed Microfibril domain |
| VWF | Proposed site of von Willebrand Factor binding |
| X-LINK | Intermolecular cross-link site |
| Orange | Cell interaction domain |
| Yellow | Keratan sulfate proteoglycan binding region on the fibril |
| Blue | Glycation region on the fibril |
| Green | Phosphorylating binding region on the fibril |
| Pink | Dermatan/chondroitin sulfate proteoglycan/decorin binding region on the fibril |
| Black Brackets | Matrix interaction domain |
| Red Brackets | Discoidin clusters of lethal OI mutations on the α2(I) chain |
| Blue Boxes | Regions containing 3 or more first position glycines on both chains where no human mutations have been reported. |
Domain Model of the Collagen Fibril

Inoglycan chains. PGs bound to the fibril may thus be expected to strongly influence fibril ligation of other factors competing for overlapping or neighboring binding sites. On the other hand, GFPGER502–507, potentially the most critical cell interaction sequence of the fibril (see below), along with the matrix metalloproteinase (MMP) 1, 2, and 13 cleavage sequence and several other key ligand binding sites, localizes to the “b1/b2” bands region where PGs are absent (Fig. 2, wide vertical non-shaded region). These initial observations suggested the collagen fibril to be organized into two domains, one mediating dynamic cell-collagen interactions, and the other carrying out structural duties. Analysis of the map and microfibril structure model provided further support for this hypothesis, as detailed below.

Cell Interaction Domain—Cell-associated molecules proposed to bind type I collagen include the integrins, PGs, discoidin domain receptors, and molecules bridging cell surfaces and the extracellular matrix, such as fibronectin and SPARC (secreted protein, acidic, and rich in cysteine) (Fig. 2). However, evidence suggests integrin binding to be the most crucial determinant in cell-collagen interactions. Integrins that bind type I collagen include the α1β1, α2β1, α1β1, and αVβ3 receptors. Because the αVβ3 integrin is thought to only interact with degraded collagen, it will not be further considered. Candidate α1β1/α2β1/α1β1 integrin binding sequences on type I collagen have been mapped to residues 127–132, 502–507, and 811–816 (Fig. 2).

GFPGER502–507: Ground Zero of Type I Collagen—Analysis of published data in the context of the map suggests a critical role for the central integrin binding site, GFPGER502–507 (Figs. 2 and 3A), according to four lines of evidence. First, this sequence binds α1β1/α2β1/α1β1 integrins and supports angiogenesis, endothelial activation, osteoblast differentiation, and cell adhesion. Second, it resides in the center of the largest “PG-clear zone” (Figs. 2 and 3A, wide vertical non-shaded region), ensuring its availability, and consistent with its dedicated function. Third, GFPGER502–507 lacks reported mutations, and also coincides with zones containing contiguous stretches of three or more first position glycines that are silent for mutations on the α1 and α2 chains (Figs. 2 and 3A, blue boxes), and in the fibril on M1 and M5 (Fig. 3B, zones 2 and 7). Such zones may be so critical to collagen function that mutations occurring there are embryonically lethal (37). These data imply that, during embryogenesis, cell-collagen interactions using GFPGER502–507 are crucial for fibril assembly, remodeling, or morphogenic processes such as angiogenesis. Moreover, considering the mutation density on the map, statistical analyses suggest that, although there is a high probability that the seven mutation silent zones may exist by random chance, it is very improbable they would align within such a narrow region of the D-period (see “Statistics” below), implying its special functional significance. Fourth, GFPGER502–507 is also a near neighbor of MLBR2, the highest concentration of ligand-binding sites of the fibril, including the site for cleavage by MMP-1, -2, and -13 required for collagen remodeling (Fig. 4A) (10). This constellation of sites also falls largely within the PG-clear zone. Analysis of the x-ray diffraction structure of the microfibril (16) confirmed these key sequences to be near neighbors, only ~3–11 nm apart on the 67 nm wide D-period (Fig. 4B). Thus, on the collagen microfibril, elements of M3 and M4 within the b1/b2 bands region, including GFPGER502–507, the MMP cleavage site, and MLBR2 are proposed to comprise a cell interaction domain where cell-collagen interactions and fibril remodeling are controlled by cell surfaces (Figs. 2 (orange box) 4).

Two assumptions regarding the importance of GFPGER502–507 to fibril function deserve further comment. First, GLPGER127–132 and GASGER811–816 in type I procollagen and in TPH form also bind α1β1/α2β1 integrins (38). At the level of the fibril, these sequences may overlap with PG binding zones, and at the monomer and fibril levels do not coincide with large mutation silent zones on both the α1 and α2 chains, distinguishing them from GFPGER502–507. Yet, the relative functions of GFPGER502–507, GLPGER127–132, and GASGER811–816 in cell-collagen interactions must be resolved through further experimental investigation. A key consideration is whether these sites prove to be available for cell interactions in the native fibril. Second, a murine knockout of the α2β1 integrin receptor yielded mice with a largely normal phenotype (39), which seems at odds with the proposal made here that GFPGER502–507 is crucial for cell-fibril interactions and embryogenesis. However, in the integrin null mouse, compensation for α2β1 function could potentially arise from the expression of other collagen-interactive receptors, including the α1β1 or α1β1 integrins.

Statistics—Statistics was used to determine whether the localization of the seven mutation silent zones to a relatively narrow fibril region (Figs. 2, 3A, and 3B) is due to random chance. Thus, several properties related to the spatial distribution of first-position glycine (G1 site) mutations in collagen were investigated. To compute the p values of interest required the derivation of some novel combinatorial quantities and algorithms. A complete presentation of these analyses appears in the supplemental materials; the following is a summary of the findings. G1 site mutations do not occur in a spatially uniform way along the collagen monomer (p values 3.6 × 10−27 and 4.3 × 10−17) under two reasonable models of spatial homogeneity. The adjacency patterns of mutation-free G1 sites, however, are quite consistent with the hypothesis that they have no tendency to cluster together or remain apart (p values all ~0.5). On the other hand, several long sequences of adjacent mutation-free G1 sites are arranged within a narrow vertical region of the D-period. The probability of their falling in this region, under a uniform random placement model, is 4.6 × 10−4. The probability of their falling in any such region throughout the entire D-period is no more than 0.0025. Thus, the apparent clustering of mutation-free zones on the fibril likely did not arise randomly.

Collagen Glycation and Ligand-Collagen Interactions—Glycation is the non-enzymatic addition of sugar to protein that occurs during aging and in an accelerated fashion in diabetes (40, 41). Thus, glucose reacts with the ε-amino group of a free lysine residue to generate a Schiff-base intermediate, followed by its rearrangement to the more stable Amadori product. Through subsequent reactions and modifications, fructosyl-lysines may then form cross-links with other proteins, forming advanced glycation end products. Collagen thus modified may become less flexible and exhibit altered cell- and ligand-colla-
gen interactions and fibrillar structure (41–45). Glycation occurs on numerous collagen residues but preferentially on hydroxylysines \( \alpha 1(I) 434 \), \( \alpha 2(I) 453, 479 \), and 924 (Fig. 2), and the “c1” band fibril region (Figs. 2 and 3A, blue stripe). Because glycated collagen is a poor substrate for cell interactions, it is notable that lysine 479 and GFPGER\(^{502-507} \) are near neighbors on the two-dimensional fibril map (Fig. 2). To examine this relationship, a three-dimensional model was built of the triple helix spanning residues 475–510, with GFPGER\(^{502-507} \) ligated to the \( \alpha 2\beta 1 \) integrin-I-domain (Fig. 3C). The lysine 479 fructosyl adduct was found to project outwards from the triple helix, poised to affect collagen-ligand interactions. Further, whereas glycation would likely not directly interfere with integrin-I-domain binding, interactions of the integrin heterodimer may potentially be affected. Examining cell and integrin interactions with native and glycated THPs, including sequences spanning the integrin binding site and lysine 479, could test such hypotheses. Lastly, it was reported that the fibril binding zone for keratan sulfate PGs, including the “c” bands, overlaps with all of the predominant glycation sites, in contrast
...potentially stabilizing its complement of key sites, or otherwise minal cross-links closely bracket the cell interaction domain, collagen in the a and c band regions. Notably, the N- and C-ter-
cross-links in type I collagen exist between residues 87 and 930, heparin binds the a bands region. Moreover, intermolecular
keratan sulfate PGs fall within the a and c bands regions, and Thus, dermatan sulfate PGs occupy the e and d bands regions,
(Figs. 2 and 4), which is supported by various lines of evidence.

...sequences of glycation to cell- and ligand-collagen interactions. Reduced affinity for glycated collagen (46) Together these
keratan sulfate PGs, but not dermatan sulfate PGs, exhibit bands (9). Consistent with this observation, it was reported that
to that of dermatan sulfate PGs that localize to the “d” and “e” bands (9). Consistent with this observation, it was reported that keratan sulfate PGs, but not dermatan sulfate PGs, exhibit reduced affinity for glycated collagen (46). Together these observations justify further investigations into the consequences of glycation to cell- and ligand-collagen interactions.

**Matrix Interaction Domain**—The map suggests that, aside from the cell interaction domain, the remainder of the D-period comprises a substrate for the binding of structural molecules (Figs. 2 and 4), which is supported by various lines of evidence. Thus, dermatan sulfate PGs occupy the e and d bands regions, keratan sulfate PGs fall within the a and c bands regions, and heparin binds the a bands region. Moreover, intermolecular cross-links in type I collagen exist between residues 87 and 930, and between type I collagen at residues 1023–1036 and type V collagen in the a and c band regions. Notably, the N- and C-terminal cross-links closely bracket the cell interaction domain, potentially stabilizing its complement of key sites, or otherwise influencing its function or availability. Next, the gap or “hole zone” and its associated proteins are proposed to contain the nucleation sites upon which hydroxyapatite crystal growth during bone mineralization occurs. Last, several domains important for collagen fibrillogenesis map to the a bands region. Thus, the overlap zone and lateral borders of the gap zone may comprise a “matrix interaction domain” where organic and inorganic matrix molecules bind to connect the fibril with other extracellular matrix elements, and to impart distinct structural properties to collagen fibrils and the tissue stroma in which they reside (Figs. 2 (black brackets) and 4).

**Non-overlapping Fibril Domains**—Integrins and MMPs (<5 nm wide) may be envisioned to interact with their binding sites in the cell interaction domain of ~20 nm wide and thus mainly influence ligands binding adjacent sequences. However, the impact of fibril-bound PGs to cell-collagen interactions may be more considerable. For example, the decorin core protein of ~40 kDa may contain a glycosaminoglycan chain averaging 40 kDa (47). The core protein is proposed to bind collagen monomer(s) in the d/e bands region. However, if its glycosaminoglycan chain is unrestrained, it could reach 80 nm or more, potentially disrupting cell- and ligand-collagen interactions at distant sites. Instead, it is proposed that the anionic glycosaminoglycans are mostly constrained within the matrix interaction domain via their binding to the transverse electropositive bands on the fibril surface, adjacent to where the PG core protein binds (Figs. 1, 2, and 5). Thus, the fibril is viewed to consist of non-overlapping cell and matrix interaction domains (Figs. 4 and 5). Nonetheless, some cross-talk between domains may be likely, e.g., if glycation affects integrin-GFPGER502–507 ligation, or matricellular proteins like SPARC with broad fibril-binding footprints influence the function of both domains simultaneously (Figs. 2 and 4).

That numerous functional sites of the type I collagen molecule fall into either of two fibril domains according to their function and that a preferred glycation substrate, lysine 479, coincides with the major fibril glycation zone support the accuracy of the overlap model of collagen fibril structure (48) upon which the map is based. Lastly, FACIT (fibril-associated collagens with interrupted triple helices) collagens (49) are proposed to interact with type I or II fibrils to influence inter-monomer

---

**FIGURE 4. Domain model of the collagen fibril.** A, schematic of domains on the collagen fibril D-period. The putative cell interaction domain, including the α1(1)α2(1) integrin binding sites GFPGER502–507 (maroon horseshoe); MMP cleavage site (green arrow); and elements of MLR2 (geometric figures); occupy the fibril overlap zone. The putative matrix interaction domain, including sites for intermolecular cross-linking (X); PG-binding (yellow and pink); and mineralization occupy the remainder of the fibril. B, physical mapping of the cell and matrix interaction domains on the collagen microfibril. Selected elements of the putative cell and matrix interaction domains were mapped, including GFPGER502–507 (maroon); MMP cleavage site (green arrow), MMP interaction domain (cyan); lateral edges of the fibronectin binding site (blue); keratan sulfate PG binding sites (yellow); dermatan sulfate PG binding sites (pink); N- and C-terminal intermolecular cross-links (black and blue arrows), respectively. The SPARC binding site (black) was rendered transparent where it overlaps the fibronectin and MMP binding/cleavage sites. The three-dimensional image was derived from an x-ray diffraction structure of the collagen microfibril in situ. The microfibrillar aspect that faces toward the outside of the fibril is at the top of the picture (the C terminus points outside of the fibril, while the N terminus faces inside the fibril). Distances between GFPGER502–507 and domains for MMP interaction, or the N-terminal end of the fibronectin-binding domain were 11.2 and 5.0 nm, respectively. A view of microfibril function in the context of the collagen fibril appears in Fig. 5.
associations within the fibril or inter-fibril dynamics (50). Whether FACIT collagens modulate the function of the cell and matrix interaction domains proposed here must await further understanding of the physical nature of FACIT-fibril interactions.

Domains in Other Fibrillar Collagens—The best characterized elements of the cell and matrix interaction domains of the human fibril as proposed here include sites for integrin binding, GFPGER502–507; MMP-1 cleavage, GPQGIA; and N- and C-terminal intermolecular cross-linking, GMKGHR and GIKGHR, respectively (Fig. 2). These elements were searched for in other fibrillar collagens (data not shown). The sequences, and their positions in the triple helices relative to the cross-link sites, are identical or highly homologous in type I collagens of other vertebrates and in the human type II/XI collagen heterotypic fibril predominant in cartilage. Thus, it may be speculated that vertebrate fibrillar collagens share the same domain structure.

Domains in Heterotypic Fibrillar Collagens—The type I collagen heterotypic fibril often contains type III and V collagens, in which some of the domain elements are altered in ways predicted to be functionally significant (51). It thus is proposed that the physical and cell-interactive properties of heterotypic fibrils may be modulated according to their contents of minor collagens carrying variants of one or more key sequences. For example, type V collagen, which lacks both an active integrin binding site at the appropriate position and an MMP-1 cleavage site, when incorporated in the type I fibril, may hinder cell interactions and fibril remodeling in tissues where low rates of collagen metabolism are desirable, such as the cornea.

Domains and the Polarity of Fibril Assembly—That type I and type II collagens and their fibrillar collagen-binding partners may share the same domain structure implies they assemble in parallel into the heterotypic fibril, where key functional domains fall into register. That atypical mutations in type I and III collagens appear at disparate regions of their triple helical domains, yet cluster to common zones of the fibril supports this contention (supplemental Fig. S1). However, although collagen monomer assembly into fibrils in vivo occurs most commonly in a head-to-tail orientation, head-to-head, or tail-to-tail assembly is also seen and plays a role in tissue morphogenesis and repair (52). It is thus notable that GFPGER502–507, proposed here to be the most critical sequence of type I collagen, is located midway between the N- and C-terminal cross-links. Therefore, if monomers assemble in a unipolar or bipolar direction into fibrils, GFPGER502–507 remains accessible for cell interactions.

Regulation of Cell-Fibril Interactions—The domain structure of the collagen microfibril was next considered in the context of the multivalent collagen fibril (Fig. 1). Although the fine structure of the collagen fibril surface remains incompletely understood, our following speculations assume that, at least in some cases, ligand-binding sites may align between microfibrils, as has been observed for some collagen-binding ligands (8, 53), although there are exceptions (54). Thus, in the fibril it is proposed that cell interaction domains appear every 5–6 nm across
Domain Model of the Collagen Fibril

the fibril’s width, and at 67 nm intervals along its length (Fig. 5). Assuming integrin heterodimers to be \( \pm 10 \) nm in diameter (55, 56) and collagen microfibril diameters \( \sim 5–6 \) nm (10), GFPGER502–507-bound integrins are proposed to cluster in stripes perpendicular to the long axis of the fibril, with one integrin bound per one or two microfibrils. According to this model, the fibril is an ideal substrate for integrin clustering, considered a key component of receptor activation and signaling (51, 56). Other collagen-binding ligands of \( \sim 5–10 \) nm in diameter may potentially interact with one or more cell interaction domains, each occupying areas of \( \sim 5 \times 20 \) nm on the fibril. It follows that the cell interaction domain may accommodate one, or at most several ligands coincidently, implying that cell- and ligand-binding interactions may be coordinately regulated. Therefore, it is speculated that integrins may positively or negatively modulate the interactions of other ligands with the fibril, or vice versa. Furthermore, MMP cleavage of one monomer may destabilize the fibril, displacing ligands, including integrins from neighboring monomers, or, conversely, could facilitate cell-fibril interactions (11) (not shown).

In summary, analysis of the distribution of functional sites and mutations on a two-dimensional model of the type I collagen fibril, and on an x-ray diffraction structure of the microfibril in situ, has identified candidate cell and matrix interaction domains. Moreover, such fibril domains may exist in heterotypic type I fibrils and other vertebrate fibrillar collagens. Defining the fine structure of the fibril domains and their functional inter-relationships may open new avenues for the therapeutic modulation of collagen function and metabolism in diseases, including fibrosis, atherosclerosis, and OI, where the protein plays a prominent role and for the engineering of fibrillar collagens and synthetic polymers for numerous applications in human medicine.

Acknowledgments—We thank Laila Huq for advice on phosphoprotein; Francisca Malfait for input regarding atypical mutations; Gregg Fields for discussion on MMPs; Richard Farndale for comments on von Willebrand Factor and discoidin domain receptor-2 collagen interactions; Sergey Leikin, Barbara Brodsky, Anton Persikov, Dale Bodian, Warren Ewens, Olena Jancenko, Donald Gullberg, and Morris Karnovsky for helpful comments; and Andrew Likens and Nanita Barchi for artwork. Please note that an interactive database containing a library of collagen ligand-binding sites and mutations integrated with physicochemical properties of the collagen proteins is under construction.

REFERENCES

1. Piez, K. A., and Reddi, A. H. (1984) Extracellular Matrix Biochemistry, Elsevier, New York
2. Ayad, S., Boot-handford, R. P., Humphries, M. I., Kadler, K. E., and Shuttleworth, C. A. (1998) The Extracellular Matrix Facts Book, 2nd Ed., Academic Press, San Diego
3. Prockop, D. J., and Kivirikko, K. I. (1995) Annu. Rev. Biochem. 64, 403–434
4. Kadler, K. E., Baldock, C., Bella, J., and Boot-Handford, R. P. (2007) J. Cell Sci. 120, 1955–1958
5. Marini, J. C., Forlino, A., Cabral, W. A., Barnes, M. A., San Antonio, J. D., Milgrom, S., Hyland, J. C., Korkko, J., Prockop, D. J., De Paepe, A., Coucke, P., Symoens, S., Glorieux, F. H., Roughley, P. J., Lund, A. M., Kuurila-
6. Svoih, K., Hartikka, H., Cohn, D. H., Krakow, D., Mottes, M., Schwarze, U., Chen, D., Yang, K., Kuslich, C., Troendle, J., Dalgleish, R., and Byers, P. H. (2007) Hum. Mutat. 28, 209–221
7. Yang, C., Hillas, P. J., Baex, J. A., Nokelainen, M., Balan, J., Tang, J., Spiro, R., and Polarek, J. W. (2004) BioDrugs 18, 103–119
8. Prockop, D. J., and Kivirikko, K. I. (1984) N. Engl. J. Med. 317, 376–386
9. Scott, J. E. (1988) Biochem. J. 252, 313–323
10. Sweeney, N. M., Di Lullo, G., Slater, S. J., Martinez, J., Iozzo, R. V., Laver-Fields, J. L., Fields, G. B., and San Antonio, J. D. (2003) J. Biol. Chem. 278, 30516–30524
11. Prockop, D. J., and Kivirikko, K. I. (1984) Proc. Natl. Acad. Sci. U. S. A. 81, 2852–2859
12. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Protetta, S., and Weiner, P. (1981) J. Am. Chem. Soc. 103, 765–784
13. Lee, K. S., Song, H. R., Cho, T. J., Kim, H. J., Lee, T. M., Jin, H. S., Park, H. Y., Kang, S., Jung, S. C., and Koo, S. K. (2006) Hum. Mutat. 27, 599
14. Ward, L. M., Lalic, L., Roughley, P. J., and Glorieux, F. H. (2001) Hum. Mutat. 17, 434
15. Palloso, D., Hart, P. S., Cortelli, J. R., Vian, S., Wright, J. T., Korkko, J., Brunoni, D., and Hart, T. C. (2001) Arch. Oral Biol. 46, 459–470
16. Malfait, F., Symoens, S., De Backer, J., Hermanss-Te, T., Sakhalin, N., Lapierre, C. M., Coucke, P., and De Paepe, A. (2007) Hum. Mutat. 28, 387–386
17. Gensure, R. C., Makitie, O., Barclay, C., Chan, C., Depalma, S. R., Bastep, M., Abuzahra, H., Cooper, R., Mundlos, S., Silence, D., Alallo Kokko, L., Seidman, J. I., Cole, W. G., and Jupan, H. (2005) J. Clin. Invest. 115, 1250–1257
18. Pollitt, R., McMahon, R., Nunn, J., Bambord, R., Alfii, A., Bishop, N., and Dalton, A. (2006) Hum. Mutat. 27, 716
19. Yoneyama, T., Kasuya, H., Onda, H., Akagawa, H., Hashiguchi, K., Naka-jima, T., Tori, T., and Inoue, I. (2004) Stroke 35, 443–448
20. Cabral, W. A., Makareeva, E., Letocho, A. D., Sibranu, F., Fertala, A., Steplewski, A., Keene, D. R., Persikov, A. V., Leikin, S., and Marini, J. C. (2007) Hum. Mutat. 28, 396–405
21. Korkko, J., Allo-Kokko, L., De Paepe, A., Nuytlinck, L., Earley, J., and Prockop, D. J. (1998) Am. J. Hum. Genet. 62, 98–110
22. Knight, C. G., Morton, L. F., Peachey, A. R., Tuckwell, D. S., Farndale, R. W., and Barnes, M. J. (2000) J. Biol. Chem. 275, 32237–32249
23. Knight, C. G., Morton, L. F., Peachey, A. R., Tuckwell, D. S., Farndale, R. W., and Barnes, M. J. (2000) J. Biol. Chem. 275, 35–40
24. Zhang, W. M., Kapyla, J., Puranen, I. S., Knight, C. G., Tiger, C. F., Pentikainen, O. T., Johnson, M. S., Farndale, R. W., Heino, J., and Gullberg, D. (2003) J. Biol. Chem. 278, 7270–7277
25. Giancotti, F. G., and Ruoslahti, E. (1999) Science 285, 1028–1032
26. Hay, E. D. (ed) (1991) Cell Biology of Extracellular Matrix, Plenum Press, New York and London
27. Persikov, A. V., Ramshaw, J. A., and Brodsky, B. (2005) J. Biol. Chem. 280, 19543–19549
28. Makareeva, E., Mertz, E. L., Kuznetsova, N. V., Sutter, M. B., DeRidder, A. M., Cabral, W. A., Barnes, M. A., McBride, D. J., Marini, J. C., and Leikin, S. (2008) J. Biol. Chem. 283, 4787–4798

3 D. L. Bodian and T. E. Klein, manuscript in preparation.
36. Byers, P. H., Wallis, G. A., and Willing, M. C. (1991) *J. Med. Genet.* 28, 433–442
37. Scott, J. E., and Tenni, R. (1997) *Cell Biochem. Funct.* 15, 283–286
38. Xu, Y., Gurusiddappa, S., Rich, R. L., Owens, R. T., Keene, D. R., Mayne, R., Hook, A., and Hook, M. (2000) *J. Biol. Chem.* 275, 38981–38989
39. Holtkotter, O., Nieswandt, B., Smyth, N., Muller, W., Schulte, V., Krieg, T., and Eckes, B. (2002) *J. Biol. Chem.* 277, 10789–10794
40. Tsilibary, E. C. (2003) *J. Pathol.* 200, 537–546
41. Paul, R. G., and Bailey, A. J. (1996) *Int. J. Biochem. Cell Biol.* 28, 1297–1310
42. Brennan, M. (1989) *J. Biol. Chem.* 264, 20947–20952
43. McCarthy, A. D., Etcheverry, S. B., Bruzzone, L., Lettieri, G., Barrio, D. A., and Cortizo, A. M. (2001) *BMC Cell Biol.* 2, 16
44. Paul, R. G., and Bailey, A. J. (1999) *Int. J. Biochem. Cell Biol.* 31, 653–660
45. Chen, J., Brodsky, S., Li, H., Hampel, D. J., Miyata, T., Weinstein, T., Gafter, U., Norman, J. T., Fine, L. G., and Goligorsky, M. S. (2001) *Am. J. Physiol.* 281, F71–F80
46. Reigle, K. L., Di Lullo, G., Turner, K. R., Last, J. A., Chervoneva, I., Birk, D. E., Funderburgh, J. L., Elrod, E., Germm, M. W., Surber, C., Sanderson, R. D., and San Antonio, J. D. (2008) *J. Cell Biochem.* in press
47. Vogel, K. G., and Heinegard, D. (1985) *J. Biol. Chem.* 260, 9298–9306
48. Chapman, J. A. (1974) *Connect. Tiss. Res.* 2, 137–150
49. Shaw, L. M., and Olsen, B. R. (1991) *Trends Biochem. Sci.* 16, 191–194
50. Eyre, D. R., Pietka, T., Weis, M. A., and Wu, J. J. (2004) *J. Biol. Chem.* 279, 2568–2574
51. Sijlander, P. R., Hamaia, S., Peachey, A. R., Slatter, D. A., Smethurst, P. A., Ouweland, W. H., Knight, C. G., and Farndale, R. W. (2004) *J. Biol. Chem.* 279, 47763–47772
52. Kadler, K. E., Holmes, D. F., Trotter, J. A., and Chapman, J. A. (1996) *Biochem. J.* 316, 1–11
53. San Antonio, J. D., Lander, A. D., Karnovsky, M. J., and Slayter, H. S. (1994) *J. Cell Biol.* 125, 1179–1188
54. Holmes, D. F., Gilpin, C. J., Baldock, C., Ziese, U., Koster, A. J., and Kadler, K. E. (2001) *Proc. Natl. Acad. Sci. U. S. A.* 98, 7307–7312
55. Nermut, M. V., Green, N. M., Eason, P., Yamada, S. S., and Yamada, K. M. (1988) *EMBO J.* 7, 4093–4099
56. Emsley, J., Knight, C. G., Farndale, R. W., Barnes, M. J., and Liddington, R. C. (2000) *Cell* 101, 47–56
57. Thomas, E. K., Nakamura, M., Wienes, D., Isacke, C. M., Pozzi, A., and Liang, P. (2005) *J. Biol. Chem.* 280, 22596–22605
58. Koide, T., Takahara, Y., Asada, S., and Nagata, K. (2002) *J. Biol. Chem.* 277, 6178–6182
59. Cabral, W. A., Makareeva, E., Colige, A., Letocha, A. D., Ty, J. M., Yeowell, H. N., Pals, G., Leikin, S., and Marini, J. C. (2005) *J. Biol. Chem.* 280, 19259–19269
60. Makareeva, E., Cabral, W. A., Marini, J. C., and Leikin, S. (2006) *J. Biol. Chem.* 281, 6463–6470
61. Morello, R., Bertin, T. K., Chen, Y., Hicks, J., Tonachini, L., Monticone, M., Castagnoli, P., Rauch, F., Glorieux, F. H., Vranka, J., Bachinger, H. P., Pace, J. M., Schwarze, U., Byers, P. H., Weis, M., Bernades, R. J., Eyre, D. R., Yao, Z., Boyce, B. F., and Lee, B. (2006) *Cell* 127, 291–304
62. Wang, H., Fertala, A., Ratnarr, B. D., Sage, E. H., and Jiang, S. (2005) *Anal. Chem.* 77, 6765–6771
63. Fujisawa, R., Zhou, H., and Kuboki, Y. (1994) *Connect. Tiss. Res.* 31, 1–10
64. Huq, N. L., Loganathan, A., Cross, K. I., Chen, Y. Y., Johnson, N. I., Willetts, M., Veith, P. D., and Reynolds, E. C. (2005) *Arch. Oral Biol.* 50, 807–819
65. Lismam, T., Raynal, N., Groeneveld, D., Maddox, B., Peachey, A. R., Huizinga, E. G., de Groot, P. G., and Farndale, R. W. (2006) *Blood* 108, 3753–3756
66. Konitsiotis, A. D., Raynal, N., Bihan, D., Hohenester, E., Farndale, R. W., and Leitinger, B. (2008) *J. Biol. Chem.* 283, 6861–6868
67. Baronas-Lowell, D., Lauer-Fields, J. L., and Fields, G. B. (2004) *J. Biol. Chem.* 279, 952–962
68. Reyes, C. D., and Garcia, A. J. (2004) *J. Biomed. Mater. Res. A* 69, 591–600