Comparative analysis of the noncollagenous NC1 domain of type IV collagen: Identification of structural features important for assembly, function, and pathogenesis

KAI-OLAF NETZER, KO SUZUKI, YOSHIFUMI ITOH, BILLY G. HUDSON,
AND RAJA G. KHALIFAH

Department of Biochemistry and Molecular Biology, University of Kansas Medical Center, Kansas City, Kansas 66160-7421

(RECEIVED September 4, 1997; ACCEPTED February 24, 1998)

Abstract

Type IV collagen $\alpha 1$ – $\alpha 6$ chains have important roles in the assembly of basement membranes and are implicated in the pathogenesis of Goodpasture syndrome, an autoimmune disorder, and Alport syndrome, a hereditary renal disease. We report comparative sequence analyses and structural predictions of the noncollagenous C-terminal globular NC1 domain (28 sequences). The inferred tree verified that type IV collagen sequences fall into two groups, $\alpha 1$ -like and $\alpha 2$ -like, and suggested that vertebrate $\alpha 3/\alpha 4$ sequences evolved before $\alpha 1/\alpha 2$ and $\alpha 5/\alpha 6$. About one fifth of NC1 residues were identified to confer either the $\alpha 1$ or $\alpha 2$ group-specificity. These residues accumulate opposite charge in subdomain B of $\alpha 1$ (positive) and $\alpha 2$ (negative) sequences and may play a role in the stoichiometric chain selection upon type IV collagen assembly. Neural network secondary structure prediction on multiple aligned sequences revealed a subdomain core structure consisting of six hydrophobic β -strands and one short α -helix with a significant hydrophobic moment. The existence of opposite charges in the α -helices may carry implications for intersubdomain interactions. The results provide a rationale for defining the epitope that binds Goodpasture autoantibodies and a framework for understanding how certain NC1 mutations may lead to Alport syndrome. A search algorithm, based entirely on amino acid properties, yielded a possible similarity of NC1 to tissue inhibitor of metalloproteinases (TIMP) and prompted an investigation of a possible functional relationship. The results indicate that NC1 preparations decrease the activity of matrix metalloproteinases 2 and 3 (MMP-2, MMP-3) toward a peptide substrate, though not to [^{14}C]-gelatin. We suggest that an ancestral NC1 may have been incorporated into type IV collagen as an evolutionarily mobile domain carrying proteinase inhibitor function.

Keywords: basement membrane; NC1; matrix metalloproteinase; phylogeny; secondary structure prediction; sequence homology; TIMP; type IV collagen

Type IV collagen, the primary scaffolding of basement membranes, has recently emerged in humans as a family of six genetically distinct $\alpha 1$ – $\alpha 6$ (IV) chains (Hudson et al., 1993a; Zhou et al., 1994). The carboxyl terminus of these α -chains is an approximately 230-residue noncollagenous NC1 domain that is of particular interest due to its crucial role in the assembly of the type IV collagen network and its involvement in renal diseases (cf. Review by Hudson et al., 1993a, and references therein). In Goodpasture syndrome, pathogenic autoantibodies are targeted mainly to the NC1 domain of the $\alpha 3$ (IV) chain, causing an immune-mediated form of glomerulonephritis and pulmonary hemorrhage. In Alport syndrome, mutations occur in the genes encoding the $\alpha 3$, $\alpha 4$, and $\alpha 5$ (IV) chains, causing progressive nephropathy and sensorineural

deafness. Several of the mutations occur within the NC1 domain, but it is not known how such mutations cause defective assembly of type IV collagen.

NC1 is presumed to specify which α -chains associate to form distinct triple-helical protomers. Moreover, it is presumed to specify which triple-helical isoforms associate end-to-end to form dimers of these protomers (Fig. 1) (Weber et al., 1984, 1988; Dölz et al., 1988; Hudson et al., 1993a; Ries et al., 1995; Kahsai et al., 1997). The discriminatory forces that govern these specific interactions are yet to be defined, but certain ones must be encoded in the structure of NC1.

Since the three-dimensional structure of NC1 is still entirely unknown (Timpl et al., 1985; Stubbs et al., 1990), we explored the NC1 domain structure in the present study using sequence-based approaches to identify regions that possibly govern the specificity for normal assembly and that may be altered in disease. We first carried out comparative analyses of 28 NC1 sequences from five

Reprint requests to: Raja G. Khalifah, Department of Biochemistry & Molecular Biology, The University of Kansas Medical Center, 3901 Rainbow Boulevard, Kansas City, Kansas 66160-7421; e-mail: rkhalifa@kumc.edu.

Fig. 1. Model of type IV collagen protomer interaction through their C-terminal NC1 domains. Light gray, $\alpha 1$ -like chains; dark gray, $\alpha 2$ -like chains. Three such chains are associated in each protomer molecule to form the collagenous triple-helix (shown truncated), and two protomers associate through the carboxy terminal NC1 domains (consisting of two homologous subdomains A and B) to form the NC1 hexamer interaction. Note the identical orientation of subdomains A and B of each globular NC1 domain within a protomer in this model, and the "cross over" interaction of subdomains from different protomers.

mammalian and five invertebrate species (Table 1) utilizing phylogenetic and multiple alignment algorithms. We then applied neural network algorithms for secondary structure prediction that took advantage of evolutionary information inherent in sequence alignments. The self-consistency of the secondary structure predictions was verified by considering the known internal homology of NC1, as it is comprised of two homologous subdomains (designated A and B) with $\approx 35\%$ sequence identity. A novel core subdomain secondary structure prediction was obtained by *jointly* considering all 56 A and B subdomain sequences.

Since sequence alignment-based search tools, such as BLAST (Altschul et al., 1990) and FASTA (Pearson & Lipman, 1988), have failed to identify homologues of type IV collagen NC1 domains, we also utilized an alternative similarity search based on a matrix of amino acid properties; this led to the discovery a possible relatedness to another extracellular matrix-associated protein, namely the tissue inhibitor of matrix metalloproteinases (TIMP). The potentially important functional consequences of such a similarity prompted us to test the hypothesis of overlapping structure and function between these two protein families and to perform experimental studies of inhibition of matrix metalloproteinases (MMP) by various NC1 structures.

Results

Phylogeny reconstruction

Type IV collagen gene evolution was reconstructed from 28 sequences (Table 1). Our analysis utilized three different models that yielded preliminary trees: distance matrix, parsimony, and maximum likelihood algorithms. A final decision between different topologies was based on: (1) whether a topology was supported by different methods, and (2) whether a topology was supported by analysis of an increasingly larger data set. The unrooted tree shown

Table 1. NC1 sequences analyzed in this study

Sequence		Accession
<i>Ascaris suum</i>	$\alpha 2$	P27393
<i>Brugia malayi</i>	$\alpha 2$	U07224
<i>C. elegans</i>	$\alpha 1$	J05067
<i>C. elegans</i>	$\alpha 2$	P17140
<i>Drosophila</i>	$\alpha 1$	M28334
Sea urchin	$\alpha 1$	L02917
Sea urchin	$\alpha 2$	S44317
Bovine	$\alpha 3$	C39419
Bovine	$\alpha 4$	S18804
Canine	$\alpha 1$	U50933
Canine	$\alpha 2$	U50934
Canine	$\alpha 3$	U50935
Canine	$\alpha 4$	U50936
Canine	$\alpha 5$	U07888
Canine	$\alpha 6$	U50937
Human	$\alpha 1$	P02462
Human	$\alpha 2$	P08572
Human	$\alpha 3$	X80031
Human	$\alpha 4$	X81053
Human	$\alpha 5$	P29400
Human	$\alpha 6$	U04845
Mouse	$\alpha 1$	P02463
Mouse	$\alpha 2$	P08122
Mouse	$\alpha 3$	S49488
Mouse	$\alpha 4$	S47279
Mouse	$\alpha 5$	S47280
Rabbit	$\alpha 1$	B61228
Rabbit	$\alpha 4$	A45137

in Figure 2 reflects the topology chosen according to the above criteria.

The tree reveals distinct clustering of sequences into $\alpha 1$ -like and $\alpha 2$ -like super groups, as well as clustering into vertebrate and invertebrate subgroups. The first and second nodes in either vertebrate cluster represent duplication events. The first nodes created precursors of $\alpha 3$ and $\alpha 1/\alpha 5$ in one branch and $\alpha 4$ and $\alpha 2/\alpha 6$ in the other, while the second nodes created precursors of contemporary $\alpha 1$ and $\alpha 5$ sequences in one branch and $\alpha 2$ and $\alpha 6$ in the other. Clustering of identically named mammalian sequences to common branches of the tree indicates a parallel evolution of the chains in different species. Since the genes encoding $\alpha 1/\alpha 2$, $\alpha 3/\alpha 4$, and $\alpha 5/\alpha 6$ are found in closely opposed pairs on different chromosomes in mammalian genomes, the two α -chain genes in one pair cannot be inherited independently and must have the same phylogenetic development. Thus the topology of any mammalian $\alpha 1$ -like sequence should mirror that of its $\alpha 2$ -like counterpart. While the topology of the $\alpha 1$ -like sequences ($\alpha 1$, $\alpha 3$, and $\alpha 5$) proved to be quite robust in the way displayed in Figure 2 (right part), the topology of the vertebrate $\alpha 2$ -like sequences was more variable. It depended on the algorithm used, and it did not always mirror the $\alpha 1$ -like topology. The paucity of $\alpha 6$ sequences in the data set may have contributed to the less robust topology in this part of the tree and probably led to the reduced bootstrap probability of 21% (Fig. 2, left part).

To reach a decision about the best topology, we performed separate analyses of the six human chains, the only subset that contained six full length NC1 chains at both DNA and protein levels. This resulted in a preferred topology with $\alpha 4$ as first taxon, fol-

Fig. 2. Unrooted phylogenetic tree of 28 type IV collagen NC1 domain sequences. Numbers in italics indicate the branch lengths reflecting results of the neighbor-joining algorithm. Small numbers close to nodes denote bootstrap probabilities (heuristic search from 100 replicates).

lowed by $\alpha 2$ and $\alpha 6$ branching off a common node, as shown in Figure 2. This tree suggests that $\alpha 3/\alpha 4$ originated before the other chains, giving them more time to diverge and explaining the greater evolutionary distance to the other chains. Interestingly, the distance of vertebrate $\alpha 2$ and $\alpha 6$ sequences to their common ancestral sequence is greater than that of the $\alpha 1$ and $\alpha 5$ sequences, indicating different mutation rates.

A noteworthy result is that the *Drosophila* sequence designated $\alpha 1$ (Cecchini et al., 1987; Blumberg et al., 1988), the only known chain in this organism, is actually clustered with the invertebrate $\alpha 2$ sequences. This was independently confirmed by the presence of $\alpha 2$ -like "formal markers" in the *Drosophila* sequence. The formal markers are invariant residues that identify $\alpha 1$ -like or $\alpha 2$ -like sequences. $\alpha 1$ -Like markers (not shown) were sequentially (but not consecutively) a Y, a 2-residue deletion, and an H, while the markers for $\alpha 2$ -like occurred as Q, G, Y, G, H, and a Q. Attempts to identify another type IV collagen sequence in *Drosophila* have not been successful so far, raising the possibility that one chain may have been lost during evolution. If true, our results indicate that this may have been an $\alpha 1$ -like *Drosophila* chain. Alternatively, the positioning of *Drosophila* close to the center point of the tree may be indicative of an α chain with hybrid structural or functional characteristics of both groups.

Primary structure analysis

A multiple alignment of the 28 NC1 domain sequences (21 vertebrate, 7 invertebrate) revealed the set of conserved or invariant

residues and the set of evolutionarily variable residues. The invariant set of 76 amino acid residues comprise about one third of the NC1 sequence and most frequently include cysteine (12), glycine (11), proline (8), serine (8), and phenylalanine (5). The highest percentages among the conserved residues are nonpolar aliphatic and aromatic residues (46%), followed by polar uncharged residues (39%) and very few charged amino acids (12%). When the two subdomains are compared in this regard, a greater number of nonpolar and charged residues occur in subdomain B. Forty-three of the invariant residues are found in the carboxyl terminal subdomain B, versus 33 in subdomain A, suggesting that there might be increased selective pressure in subdomain B to maintain residues with possible structural or functional importance.

Further comparison of subdomains A and B reveals that there is a subgroup of invariant residues that is conserved not only across species, but also across subdomains. This subgroup (marked by asterisks in Fig. 6 below) comprises some 20 residues which occur in the order: HSQ,P-C,G-S,Q,GSC,F—P,C,C,W,P,RC-VC, where the dashes indicate single intervening residues and the commas separate longer intervals. The subdomain conservation of these residues, which prominently include the cysteines that form the set of three disulfides, dates back to a time prior to the duplication event that first created an NC1 molecule consisting of two homologous subdomains. It must be inferred that these residues are the most essential toward maintaining structure in each subdomain of NC1.

In contrast to the above, the variable residues in NC1 presumably confer some structural and functional specificity to the individual NC1 domain. We analyzed the variability at every single

position in the alignment (cf. Sander & Schneider, 1991), and positions with scores greater than 10% were identified. Fifty-three residues in subdomain A, and 56 residues in subdomain B belong to this group. Clusters of variable residues are found at the N- and C-terminal ends, around the first, and between the fourth and fifth cysteine residues of each subdomain (vide infra).

To identify which of the variable residues possibly confer group-specific ($\alpha 1$ -like or $\alpha 2$ -like) functionality, "average" consensus sequences were extracted from $\alpha 1$ -like and $\alpha 2$ -like sequence profiles generated by the program PROFILEMAKE (GCG). The two consensus sequences were then compared by using the BESTFIT program (GCG). This procedure yielded a set of about 25 residues per subdomain in which the $\alpha 1$ -like and $\alpha 2$ -like consensus sequences differed significantly, as indicated by thick dark bars in Figure 3. In this set of residues, the fraction of polar and charged amino acid residues ($\alpha 1$ -like 65%, $\alpha 2$ -like 69%) is appreciably higher than among the group of conserved or invariant residues (57%). It is notable that opposite charges are conferred by these group-specific residues, particularly in the B subdomain, with an excess of positive charge in the $\alpha 1$ -like (+3) and a striking excess of negative charge in the $\alpha 2$ -like (-4) sequences.

Secondary structure prediction

We derived a 3-state (α -helix, β -strand, random coil) secondary structure prediction from multiple NC1 sequence alignments by using the neural network algorithm PHD of Rost and Sander (1993b). The availability of a large set of NC1 sequences spanning a wide range of sequence identity (54–98%) made such an analysis particularly propitious. Figure 4A displays the PHD-predicted (multiple-sequence derived) probabilities for β -strand and α -helix superimposed over the full length of the human $\alpha 1$ (IV) NC1 sequence. Visual inspection of the β -strand probability revealed a high degree of similarity between subdomains A and B in our PHD prediction, which is not normally seen in previous single-chain predictions. More objec-

tively, an auto-correlation analysis of the PHD β -strand prediction yielded the highest correlation coefficient (0.309) for a duplication with a lag number of 112 residues. This internal homology in β -strand prediction begins immediately after the sixth cysteine residue, i.e., the beginning of subdomain B, in perfect match with the primary sequence. A similar maximum in the autocorrelation was obtained by utilizing the combined probabilities for β -strand, α -helix and solvent accessibility (data not shown).

The PHD prediction method was also applied to shed light on the question of possible conformational differences between different classes of α -chains, but it did not reliably identify differences between the $\alpha 1$ -like and the $\alpha 2$ -like NC1 sequences. If real differences exist, they would be more difficult to detect due to the reduced diversity of the input set of sequences when they were divided into $\alpha 1$ -like and the $\alpha 2$ -like groups. In reverse of the above, a statistically more reliable secondary structure prediction was obtained when the input to the sequence alignment was the unified set of all 56 A and B subdomain sequences (two from each NC1 domain). The increased evolutionary information content (sequence identity range now 30–98%) led to higher probabilities for two of the predicted β -strands as well as decreased overlap of the probability curves for α -helix (dashed curve) and β -strand (solid curve), as shown in Figure 4B. The α -helix prediction (peak labeled H1) was also strengthened.

The combined subdomain prediction arguably represents the most likely common core structure of NC1 subdomains, which appears to consist of six β -strands (35% of residues) labeled E1–E6, and one α -helix (<5%). Figure 4B also plots the PHD-predicted solvent accessibility, hydrophilicity, and sequence variability based on this comparison of 56 NC1 subdomain sequences. The predicted β -strands for the most part are clearly seen to occur in sequences predicted to have little or no solvent accessibility and decreased hydrophilicity. There was no distinct pattern of adjacent alternating polar and nonpolar amino acids among the predicted β -strands. Most of the predicted random coils fall onto

Fig. 3. Comparison of the $\alpha 1$ -like and the $\alpha 2$ -like consensus sequences: identification of group-specific residues. Thick vertical bars denote dissimilar residues or deletions which create group-specific differences. Charges among these residues are shown by (+) or (-) symbols above the sequence. Note the predominance of negative charge in $\alpha 2$ -like subdomain B (net charge -4), which is opposed by positive charge in the same region of the $\alpha 1$ -like sequence (net charge +3). Predicted secondary structure elements (PHD) are indicated by open boxes and labeled: E, predicted β -strand; H, predicted α -helix. Cysteine residues are marked by arrowheads.

Fig. 4. Evaluation of the secondary structure prediction for NC1. **A:** Prediction for full-length NC1. Alignment of probability graphs for β -strand (top) and α -helix (bottom) predictions, as derived from the PHD algorithm. Visual inspection of the PHD β -strand prediction reveals the internal homology of NC1. The border between subdomains A and B are marked by the bold dashed line. Cysteine residues and disulfide bonding pattern are indicated by arrowheads at the top of the graph. **B:** PHD prediction for aligned NC1 subdomains (56 sequences). Top to bottom: α -helix/ β -strand probabilities, estimated solvent accessibility, hydrophilicity, and sequence variability. Predicted β -strands are labeled E1–E6, predicted α -helix is labeled H1. Note the decreased accessibility of the predicted β -strands. Invariant residues across both subdomains are shown at the bottom in the rectangular box. **C:** Helical wheel representation of the predicted α -helix in NC1 subdomains. The respective homologous sites in subdomains A and B of the human $\alpha 1(\text{IV})$ NC1 sequence are shown. Note that in both cases the predicted α -helix has a hydrophobic and a hydrophilic side (hydrophobic moment). The charge difference between the subdomains is present in almost all NC1 sequences analyzed.

areas more exposed to solvent and with increased hydrophilicity. In general, most predicted secondary structure elements fall onto regions with lower sequence variability, consistent with the evolutionary importance of these ordered structures.

A closer examination of the subdomain predictions for the putative α -helix reveals it to have both exposed and nonexposed segments. When drawn as helical wheels (Fig. 4C), the respective residues from various α -chains have a significant hydrophobic moment, indicating they might be exposed to solvent on one side but not the other. Significantly, the exposed helical regions carry *opposite charges* in subdomains A (positive, except neutral for the $\alpha 4(\text{IV})$ sequences) and B (negative). This is highly suggestive of a possible site for the interaction of oppositely charged α -helices, either within a given NC1 domain or across different NC1 domains, e.g., as might occur in the assembly of NC1 hexamers.

Similarity search for proteins related to NC1

NC1 of type IV collagen molecules bears little or no significant resemblance to pro-peptides or NC domains of other collagens

(Oberbumer et al., 1985; Dion & Myers, 1987), and earlier studies using sequence alignment-based search tools (BLAST, FASTA) have failed to identify any homologues in databases of sequenced proteins or genes. We consequently explored other approaches that might provide starting points for identifying and investigating any possible relatedness of NC1 to different proteins. The available NC1 domain sequences from vertebrate and invertebrate species (Table 1) were, thus, submitted to the EMBL-Heidelberg PROPSearch server (Hobohm & Sander, 1995) for a "similarity" search against SwissProt and PIR databases based solely on amino acid composition-related properties (for a detailed list of the 144 properties and the evaluation of parameters see Hobohm & Sander, 1995). The search was significantly improved by employing the *family search* option, thus providing evolutionary information to the algorithm. The output, summarized in Table 2, was a list of rank-ordered proteins (closest match on top) with possible "relationship" to the query sequence.

Besides the expected homologous NC1 sequences, members of the tissue inhibitor of metalloproteinases (TIMP) family, mostly TIMP-1 from various species, unexpectedly appeared in top rank-

Table 2. PROPSEARCH (Hobohm & Sander, 1995) results using the "family search" option for NC1^a

Rank	ID ^b	DIST ^c	LEN ^d	pI	Comments
1	A39786	6.13	218	7.84	Collagen alpha 3(IV) NC1 (human, fragment)
2	S47278	7.98	246	8.53	Collagen alpha 3(IV) NC1 (mouse)
3	P35624	8.05	207	7.85	Metalloproteinase inh. prec. TIMP-1 (pig)
4	P20414	8.33	207	8	Metalloproteinase inh. prec. TIMP-1 (bov)
5	S47280	8.34	253	7.51	Collagen alpha 5(IV) NC1 (mouse)
7	P01733	9.4	135	6.72	T-cell receptor beta-chain prec. V region
8	P01033	9.57	207	8	Metalloproteinase inh. prec. TIMP-1 (human)
9	B61228	9.74	258	7.71	Collagen alpha 1(IV) NC1 (rabbit, fragment)
10	P20614	10.03	206	7.69	Metalloproteinase inh. prec. TIMP-1 (rabbit)
11	P33589	10.14	228	6.9	Gyroxin analogue (EC 3.4.21.74) (thrombin-like enzyme)
13	P30120	10.25	217	8	Metalloproteinase inh. prec. TIMP-1 (rat)
15	P12032	10.48	205	8.83	Metalloproteinase inh. prec. TIMP-1 (mouse)
16	P03252	10.59	204	8.63	Thiol endoprotease (EC 3.4.22.-) (human adenovirus type 2)
20	JC4131	10.74	219	7.73	Glioma pathogenesis-related protein—human

^aMultiple listings omitted.^bID, database accession number.^cDIST, Euclidean distance score.^dLEN, length of the identified sequence.

ing positions. The average distance score for the NC1 sequences in Table 2 was 8.05, while the score for the TIMP-1 sequences was 9.45. These values are associated with >90% and >80% probabilities, respectively, that the sequences are related to the query sequence according to the criteria of Hobohm and Sander (1995). The similarities of NC1 and TIMP-1 detected by the PROPSEARCH algorithm reflect properties such as sequence lengths (≈ 230 versus 184 residues), number of cysteine residues (12 in both proteins), pI values (6–9 vs. 7–8), single amino acid occurrences (aromatic, 10.8 vs. 10.3%; H, 3.0 vs. 3.3%; G, 7.8 vs. 6.5%; L, 7.3 vs. 9.8%; Y, 3.9 vs. 3.3%), etc.

Although a structural similarity at the folding level between NC1 and TIMP is not implied by the above findings and would be an extremely remote possibility, we nevertheless explored the PHD secondary structure prediction method as a tool for comparison of these two proteins. A PHD prediction for the human TIMP-1 "core structure" was generated from an alignment of 15 homologous TIMP-1, TIMP-2, and TIMP-3 with identities in the range of 45%. The predicted secondary structure (not shown) is dominated by β -strand (34%), with α -helix accounting for only 11%. A cross-correlation analysis was carried out, with the probabilities for β -strand, α -helix, and solvent accessibility of the core NC1 subdomain being the probe vector. The best possible alignment of the NC1 β -strand prediction (the dominant secondary structure element) was found starting at position 70 of the TIMP sequence. The inner 2 disulfides of NC1 match reasonably well with TIMP, but the alternating pattern of hydrophilic and hydrophobic residues, noted as one characteristic feature of the β -strand elements in the proposed TIMP-2 β -barrel structure (Williamson et al., 1994), is absent in the predicted NC1 β -strands. (Subsequent to the submission of this manuscript, the preliminary crystallographic structure of TIMP-1 in its complex with MMP-3 appeared (Gomis-Rüth et al., 1997), as discussed in the Addendum below.)

NC1 Inhibition of MMP activity

The potential functional consequences of a "similarity" between these two extracellular matrix proteins led us to investigate whether

NC1, by functional analogy to TIMP, has an inhibitory effect on matrix metalloproteinases (MMP). Figure 5 shows that MMP-2 activity is, indeed, inhibited in the presence of bovine lens basement membrane NC1 hexamer and monomer preparations (a mixture of $\alpha 1$ and $\alpha 2$) (Timoneda et al., 1990). It is apparent that NC1 seems to be more inhibitory as monomer than as hexamer. However, Figure 5 also shows that the inhibitory effect is not complete even at the highest NC1 concentrations tested (approximately 70% inhibition at 250 $\mu\text{g/mL}$). Individual $\alpha 1$ and $\alpha 2(IV)$ NC1 chains were also isolated and tested individually. The inhibitory activities of purified $\alpha 1$ and $\alpha 2(IV)$ NC1 monomers were not significantly different from each other or from the mixed monomer preparation used previously (data not shown).

Fig. 5. Inhibition of MMP-2 by NC1. MMP-2 activity was assayed by using a fluorescent peptide as substrate. Uninhibited activity was set to 100%. MMP-2 was pre-incubated with increasing concentrations of NC1 hexamer (filled triangles) and monomer (filled circles) from bovine lens basement membrane, containing a mix of $\alpha 1$ and $\alpha 2(IV)$ chains, or with TIMP-1 (open squares). Residual activity of MMP-2 is plotted versus the concentration of inhibitors.

The effect of NC1 on MMP-3 (stromelysin 1) was also tested. MMP-3 is distinguished from MMP-2 by the absence of a fibronectin-like domain, leading to unique proximity of the catalytic and the C-terminal hemopexin-like domains (Birkendal-Hansen et al., 1993; Nagase, 1996). The substrate specificity is also different from that of MMP-2 (cf. Nagase & Fields, 1996). In an experimental assay similar to that described above for MMP-2, MMP-3 showed a maximum 75% inhibition in the presence of 125 $\mu\text{g/mL}$ NC1 monomer ($\alpha 1(\text{IV})/\alpha 2(\text{IV})$ mix).

Negative control experiments ruled out the possibility that NC1 was contaminated with proteinases or that NC1 quenched substrate fluorescence, the parameter used to follow proteolysis of the synthetic peptide substrate. Silver stained protein gels after 24 h incubation showed that NC1 itself was resistant to MMP-2 and MMP-3 digestion. Thus the inhibition was less likely to be due to NC1 acting as a competitive substrate. Surprisingly, preliminary experiments using ^{14}C -labeled gelatin as macromolecular substrate for MMP-2 and denatured ^{14}C -labeled transferrin as substrate for MMP-3 did not reveal a significant inhibitory activity of NC1 in the hydrolysis of these substrates.

Discussion

The studies described above were initiated in order to gain insights into the structure and function of type IV collagen NC1 domains, their assembly, and their role in various diseases that have been linked to basement membranes. Little progress so far has been reported toward the direct elucidation of the three-dimensional structure of the NC1 domain (Timpl et al., 1985; Stubbs et al., 1990). The molecular mass of each NC1 domain (≈ 26 kDa) and its aggregation into hexamers makes high resolution solution studies by techniques such as NMR quite difficult. Consequently, we explored different comparative sequence analyses including phylogenetic tree inference, primary sequence alignment and comparison, secondary structure prediction, and searching for related proteins. These approaches were all considerably enhanced by the availability of multiple NC1 sequences spanning a wide range of homology.

The evolution of human type IV collagen NC1 has previously been studied on the basis of sequence similarities (Leinonen et al., 1994; Zhou et al., 1994), but a more formal inference of a phylogenetic tree has been reported only for the human $\alpha 1$ - $\alpha 4$ (IV) NC1 sequences (Quinones et al., 1992). The extended picture of type IV collagen gene evolution we have derived is consistent with previous, though less extensive, analyses in that all NC1 sequences belong to either of two basic groups, an $\alpha 1$ -like or an $\alpha 2$ -like group. Two of the invertebrate and all of the vertebrate species in the study contain $\alpha 1$ -like and $\alpha 2$ -like type IV collagen chains. They must have originated by duplication before the evolution of current species. Our analysis suggests that $\alpha 3/\alpha 4$ (IV) chains were the first to take an independent evolutionary development, while the lines of $\alpha 1/\alpha 2$ and $\alpha 5/\alpha 6$ (IV) separated more recently. This pattern explains the greater distance of the $\alpha 3/\alpha 4$ (IV) chains to the other chains.

It has remained an unsolved question as to which residues confer the group-specific $\alpha 1$ -like or $\alpha 2$ -like functionality (Mariyama et al., 1992; Zhou et al., 1994). We have identified group-specific variable residues based on a comparison of consensus sequences of $\alpha 1$ -like and $\alpha 2$ -like sequences that we believe contain information about the "average" $\alpha 1$ or $\alpha 2$ sequence (Fig. 3). The subset of $\alpha 1$ or $\alpha 2$ -specific residues comprised about 25 residues per NC1 subdomain with an approximately 70% fraction being polar and charged

residues. The locations of these residues coincide for some part with variable NC1 regions identified by Mariyama et al. (1992), but group-specific residues also occur before the first cysteine residue in each subdomain and at the carboxyl terminal end.

The comparison of differences among NC1 sequences becomes more meaningful when placed in the context of the structure of NC1. In absence of direct structural information, we can only consider the predicted secondary structure, which is quite consistent with the low helical content of NC1 suggested by circular dichroism measurements (Wieslander et al., 1985). Figure 6 summarizes our findings on the secondary structure elements predicted and the residue variability, as superimposed on a diagram of the consensus NC1 sequence. Most of the characteristic $\alpha 1$ -like and $\alpha 2$ -like residues fall into regions where no ordered secondary structure is predicted (open circles in Fig. 6).

Previous predictions of NC1 secondary structure were invariably based on single sequences and did not lead to detecting internal homology in the predictions or to identifying of a common core structure of subdomains (Dion & Myers, 1987; Myers et al., 1987; Kefalides et al., 1993; Matsukura et al., 1993; Monboisse et al., 1994; Ninomiya et al., 1995). Any reliable prediction should show internal consistency by predicting similar structures for the

Fig. 6. Schematic representation of a type IV collagen NC1 domain with predicted secondary structure (α -helix, H, and β -strand, E, only). The sequence shown is the consensus of 28 NC1 domain sequences. Invariant residues which are conserved among all sequences are indicated by filled circles. The (*) indicate 21 residues that are invariant in alignments of 56 subdomains.

two homologous subdomains of NC1, and our analysis successfully does so for the dominant β -sheets. Our predicted subdomain "core" structure consists mostly of hydrophobic β -strands that do not show an alternating pattern of hydrophobic/hydrophilic residues. The hydrophobic β -structure may thus build a tightly packed core that is surrounded by a surface layer of more hydrophilic, variable residues in non-ordered "loop" conformation, representing the interface to potential binding partners. A protected core structure may provide a rationalization of some of the biochemical properties of NC1, such as its remarkable resistance to denaturation *in vitro* (Wieslander et al., 1985) and to enzymatic degradation *in vivo* (Weber et al., 1984).

Our novel prediction of oppositely charged α -helices, located around the second cysteine residue in each subdomain, suggests that they could play an important role in the intra- and intermolecular interaction of NC1 domains. Siebold et al. (1988) showed that these cysteine residues are involved in a disulfide shuffling leading to cross-linking of NC1 domains from two different type IV collagen protomers. The occurrence of opposite charges in the predicted α -helices may then be involved in this "cross over" interaction of subdomain A from one NC1 domain with subdomain B from another NC1 domain. The model shown in Figure 1 would allow this cross over interaction to take place and further illustrates that, within one protomer, like subdomains would be found in the same orientation. The presence of identical charges in the disulfide exchange regions of such aligned subdomains might explain another observation of Siebold et al. (1988), namely that NC1 disulfide bonding within one protomer does not occur. Site-directed mutagenesis experiments on replacing these charges will be very valuable in determining whether subdomain A-subdomain B helix interactions are important for the disulfide exchange between NC1 domains during hexamer assembly.

The analyses presented here also provide a rational basis to design experiments aimed at elucidating the structure of the epitope that is targeted by autoantibodies in Goodpasture syndrome kidney disease. Despite homology to other NC1 domains, the $\alpha 3(\text{IV})$ NC1 domain is the primary antigen (Saus et al., 1988; Hudson et al., 1993b), and 85% of Goodpasture syndrome patient sera react

exclusively with the $\alpha 3(\text{IV})$ NC1 domain (Kalluri et al., 1995). Thus, the antigenic epitope primarily resides within the $\alpha 3(\text{IV})$ NC1 domain, and there is evidence that it is a discontinuous epitope (Kalluri et al., 1996). Our comparison of the human $\alpha 3(\text{IV})$ NC1 sequence to a consensus sequence of its closest relatives, i.e., all the $\alpha 1$ -like sequences excluding the $\alpha 3(\text{IV})$ from the data set, has identified six regions of variable residues that are candidates for the Goodpasture epitope: residues T2-T5, A18-S31, A89-E104 in subdomain A, and residues T127-K141, F148-T156, and N189-I221 in subdomain B $\alpha 3(\text{IV})$. Two regions involving cysteine residues (residues A18-S31 and T127-K141) are especially likely candidates, since the reactivity of the Goodpasture epitope is known to depend on intact disulfide bonds (Wieslander et al., 1984; Kalluri et al., 1991).

The theoretical evaluation of NC1 structural properties presented in this study also provides new knowledge of why certain mutations in type IV collagen cause nephritis, deafness, and ocular changes in patients with Alport syndrome. Table 3 summarizes amino acid substitutions found in the NC1 domain of $\alpha 5(\text{IV})$ collagen with respect to their locations and resulting phenotypes. The mutations affect predominantly invariant residues (6 of 9) conserved across species and subdomains, and they result in more severe phenotypes when they do so. Mutations that affect nonconserved residues may create drastic changes by introducing charged amino acids or disulfide forming residues, yet seem to be associated with milder phenotypes. In contrast, there is seemingly no clear connection between affected predicted secondary structure and the phenotype of the disease.

The retrieval of tissue inhibitor of metalloproteinases (TIMP)-1 was the most striking and unexpected result of the amino acid property-based similarity search. TIMPs are thought to play an important role in the control of matrix turnover by regulating matrix metalloproteinase (MMP) activity. We have compared the predicted NC1 with the predicted TIMP-1 structure in order to detect similarities at the level of secondary structure, and we do find that the predicted overall content of β -strands is similar. However, this approach cannot provide definitive evidence of folding similarities. Note that while both proteins consist of two subdomains, those of

Table 3. Mutations in Alport syndrome: Amino acid substitutions in $\alpha 5(\text{IV})$ NC1

Mutation	Invariant residue ^a	Pred. Sec. structure ^b	Phenotype	Source
Subdomain A				
Ala1498 → Asp	No	L	Mild ^c	Tverskaya et al. (1996)
Pro1517 → Thr	Yes* ^d	L, E3	Severe ^e	Lemmink et al. (1993)
Trp1538 → Ser	Yes*	E5	Mild	Lemmink et al. (1993)
Arg1563 → Gln	Yes*	E6	Severe	Zhou et al. (1993)
Cys1564 → Ser	Yes*	E6	Severe	Zhou et al. (1991)
Cys1567 → Arg	Yes*	E6	Severe	Knebelmann et al. (1996)
Subdomain B				
Gly1596 → Asp	Yes*	E2	Severe	Renieri et al. (1996)
Tyr1597 → Cys	No	E2	Mild	Netzer et al. (1996)
Leu1649 → Arg	No	E5	Mild	Barker et al. (1996)

^aInvariant residues are noted in Figure 6.

^bSecondary structure elements (E = beta strand, H = helix, L = random coil) are depicted in Figures 3 and 6.

^cMild, age at end stage renal failure >30 years.

^d* denotes invariant residue conserved across species and subdomains.

^eSevere, age at end stage renal failure <30 years.

TIMP are not internally homologous and are quite different in length (NC1, $\approx$ 110/120 residues; TIMP, 127/57 residues). Furthermore, the location of the cysteine residues and the arrangement of disulfides reveal no strong resemblance between NC1 and TIMP.

The possibility of a TIMP-NC1 relationship, however, led us to investigate whether NC1, by analogy to TIMP, has an inhibitory effect on matrix metalloproteinases. Our results show that monomeric NC1 domains, in micromolar concentrations, decreased MMP-2 (gelatinase A) and MMP-3 (stromelysin 1) activity towards a small synthetic peptide substrate (Knight et al., 1992). However, the inhibition unexpectedly was not apparent with macromolecular substrates such as ^{14}C -gelatin and ^{14}C -transferrin (denatured). It is known that MMP specificity toward small peptide substrates can differ from that toward native macromolecular substrates of (Nagase & Fields, 1996). The particular basis for the substrate-dependent differences in inhibitory effects of NC1 is not clear and remains to be investigated. It is possible that NC1 may bind to the gelatin (Tsilibary et al., 1988) or denatured transferrin substrates, thus weakening its effect on MMP-2 and MMP-3. It is also possible that the interaction of NC1 with MMP-2 or MMP-3 might be weakened or abolished by the interaction of gelatin or transferrin substrates with the enzymes. Complex interactions of NC1 have been reported with gelatin and other matrix components (Allan et al., 1995). Inhibition of both MMP-2 and MMP-3 in peptide assays argues against an involvement of the fibronectin-like domains which are unique to MMP-2. Our results suggest that NC1 binding by MMP, if confirmed, will be quite different from that of TIMP.

Regardless of the existence of any NC1-TIMP structural similarities, our results raise the important question of whether NC1 can play a role in the regulation of MMP-mediated proteolytic degradation of type IV collagen. We note that the so-called C5 domain at the carboxyl-terminus of the $\alpha 3$ chain of type VI collagen has been found to have structural homology with the Kunitz-type inhibitor of serine proteinases (Chu et al., 1990; Arnoux et al., 1995). Proteinase inhibitor functionality of the $\alpha 3(\text{VI})$ Kunitz-type domain has been lost (Arnoux et al., 1995), but it is resistant even to high concentrations of trypsin, as is type IV collagen NC1 to digestion with MMP-2. Our results raise the possibility that the ancestral NC1 sequence was incorporated into the type IV collagen precursor as an evolutionarily mobile domain (Bork, 1992) carrying proteinase inhibitor function, a similar evolutionary principle to the $\alpha 3(\text{VI})$ collagen chain.

Addendum

Following the completion of our study and the submission of this manuscript, the TIMP-1 crystallographic structure was reported for its complex with MMP-3 by W. Bode and colleagues (Gomis-Rüth et al., 1997). Unpublished secondary structure assignments for TIMP were then generously made available to us by Dr. W. Bode (pers. comm.). Our preliminary examination showed that the PHD prediction was robust for β -strands and quite poor for α -helices of TIMP. Comparison of NC1 with TIMP by approaches such as threading or 3D-profiles awaits release of the full coordinates of TIMP, as does the determination of whether a TIMP template may accommodate disulfide pairings of an NC1 subdomain.

Materials and methods

Sequence retrieval and computations

The human $\alpha 1(\text{IV})$ NC1 domain sequence was used as a query sequence for a BLASTP search of GenBank, PIR, and SwissProt

databases. Twenty-one vertebrate and seven invertebrate NC1 sequences were retrieved (Table 1). Sequence and structural computations were carried out on Macintosh (PHYLIP version 3.5; Felsenstein, 1989, 1996) or Silicon Graphics IRIS (Wisconsin Genetics Computer Group, Inc., GCG version 8.0) workstations, or via internet servers (EMBL Heidelberg server). Autocorrelation computations were programmed in MathCad worksheets (MathSoft, Inc.).

Phylogenetic tree reconstruction

Multiple alignments of DNA and protein sequences were generated by using the PILEUP program (GCG). DNA alignments were edited for correct positioning of gaps. BOOTSTRAP program (PHYLIP) was used to create 100 heuristic search replicates of the sequence alignments on which PARSIMONY program (PHYLIP) analysis was performed to find the most parsimonious tree. Distance matrices were created from bootstrap replicates with corrections for multiple substitutions (Kimura, 1980; Felsenstein, 1996). Phylogenetic trees were then deduced by using the GROWTREE program (GCG), implementing the neighbor-joining algorithm (Saitou & Nei, 1987). The CONSENSE program (PHYLIP) was applied to generate a consensus tree that incorporated a topology found in the majority of bootstrap replicates. Human sequence subsets were analyzed further by using maximum likelihood (PHYLIP) and parsimony algorithms with randomized order of taxon joining and branch-swapping.

Primary structure analysis

Sequences of the NC1 domains were extracted from the retrieved type IV collagen sequences. The beginning of NC1 was defined as the first amino acid of the most C-terminal interruption of the collagenous G-X-Y repeat. NC1 subdomain sequences were generated by separating immediately after the sixth cysteine residue to give rise to an N-terminal subdomain A ($\approx$ 110 amino acid residues) and a C-terminal subdomain B ($\approx$ 120 residues). Multiple sequence alignments were generated by PILEUP (GCG). Consensus sequences for the $\alpha 1$ -like and the $\alpha 2$ -like groups, as well as for the complete study set of 28 sequences and 56 subdomains were generated by using the program PROFILEMAKE (GCG). $\alpha 1$ -Like and the $\alpha 2$ -like consensus sequences were compared by using the program BESTFIT (GCG).

Secondary structure prediction

The PHD secondary structure prediction e-mail server at EMBL was used for secondary structure prediction (Rost & Sander, 1993a). The program additionally provides estimates for the accessibility of each residue, as well as numerical values for the probability of each residue to be in one of the predicted secondary structure states. These values could be plotted over the length of the sequence and were used as input values for auto- and cross-correlation analysis.

Correlation analysis

Probabilities for each predicted secondary structure state (H , E , L), and residue accessibility predictions derived from PHD analysis of full-length NC1 sequences were examined by autocorrelation (Zimmerman et al., 1968; Jones, 1975; Kubota et al., 1981). Without

explicit use of internal sequence homology information, autocorrelation analysis provided a test for the presence of a periodicity or a repeat in the prediction. The best position for an internal alignment was determined by correlating the probabilities for any given secondary structure type (H , E , L) or for the accessibility (A) at the k^{th} residue with those at $j + k$, where j is the lag spacing, and determining which lag value j gives the highest correlation coefficient. The autocorrelation function for the PHD prediction of beta sheet probabilities E is given by

$$\text{Autocorr}(E)_j = \frac{\sum_k^{N-j} (E_k - \bar{E}) \cdot (E_{k+j} - \bar{E})}{\sum_k (E_k - \bar{E})^2} \quad (1)$$

where the individual probabilities E_j are first computed as differences from the mean values in order to restrict the variation of the autocorrelation coefficient to the normalized range of -1 to $+1$. Cross-correlation analysis of two sequences was carried out in analogous fashion.

Search for proteins similar in properties to the NC1 domain

Single complete NC1 sequences were submitted to the PROPSearch server (Hobohm & Sander, 1995). The search algorithm converts the submitted sequence to a query vector composed of 144 weighted amino acid properties, including molecular weight, content of bulky residues, content of small residues, average charge and others (Hobohm & Sander, 1995). Sequences in the reference databases (PIR and SwissProt) are transformed accordingly to enable computation of Euclidean distances between query and database sequences. Distance scores of returned proteins were examined to determine the likelihood of a relationship between the query sequence and the closest matches.

Preparation of NC1 protein

NC1 was isolated from bovine lens basement membrane (LBM) by collagenase digestion, anion exchange and size exclusion chromatography as described (Wieslander et al., 1985). To purify $\alpha 1$ and $\alpha 2$ NC1 monomers under nondenaturing conditions, we followed the chromatofocusing procedure in Timoneda et al. (1990). Peaks containing $\alpha 1$ and $\alpha 2$ monomers were collected separately, pooled and concentrated. Hexamer and monomer concentration were adjusted to 400 $\mu\text{g/mL}$ in TBS (150 mM NaCl, 50 mM Tris-HCl pH 7.5), and the integrity of the preparations was confirmed by SDS-PAGE.

MMP proteolytic assays

Purified human pro-MMP-2 (progelatinase A) and human TIMP-1 were kindly provided by Prof. H. Nagase.

Small peptide substrate assay

Pro-MMP-2 was activated by incubation with APMA for 30 min in TNC buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10 mM CaCl_2 , 0.05% Brij) (Itoh et al., 1995). Activated enzyme was pre-incubated at 0.5 $\mu\text{g/mL}$ for 1 h at room temperature with TIMP-1 (0.5 $\mu\text{g/mL}$), NC1 (1–400 $\mu\text{g/mL}$), and bovine serum albumin (1–400 $\mu\text{g/mL}$) or buffer as controls. Undenatured NC

hexamer and monomer containing a mix of $\alpha 1$ and $\alpha 2$ (IV) chains, as well as purified $\alpha 1$ and $\alpha 2$ monomers from LBM were used. An equal volume of the peptide substrate (7-methoxycoumarin-4-yl)acetyl-Pro-Leu-Gly-Leu(2,4-dinitrophenyl)-L-2,3-diaminopropionyl-Ala-Arg-NH₂ was added to the enzyme/inhibitor reaction mix. The assay was carried out according to Knight et al. (1992). Final reagent concentrations were: MMP-2, 3.6 nM; TIMP-1, 8.4 nM; NC1, 19 nM–3.8 μM as monomer; BSA, 12.5 nM–5 μM ; and substrate, 5 mM. After 30 min incubation at 37 °C, the reaction was stopped with 1 mL 3% acetic acid, and fluorescence was measured (λ_{ex} 325 nm, λ_{em} 393 nm) on a Shimadzu F300 spectrofluorometer. Mixtures of buffer and substrate served as control (baseline fluorescence). Other controls included single components with substrate as well as the addition of NC1 after stopping the reaction in order to rule out quenching.

Protein substrate assay

Activated MMP-2 (0.2 $\mu\text{g/mL}$ final concentration) was pre-incubated with LBM NC1 hexamer, NC1 monomer, and purified $\alpha 1$ or $\alpha 2$ monomer (all at 150 $\mu\text{g/mL}$ final concentration) for 1 h at room temperature in TNC buffer before the addition of ^{14}C -labeled gelatin substrate (final concentration 1.5 mg/mL). ^{14}C nonprecipitable disintegrations per minute (dpm) were counted after 1 h in a scintillation counter (Beckman). MMP-2 was assayed with BSA (200 $\mu\text{g/mL}$) or TIMP-1 (0.2 $\mu\text{g/mL}$) for control.

Acknowledgments

This work was supported by N.I.H. grant DK 18381 to B.G.H., by DFG grant Ne 350/7-1 and grants from the Fritz Thyssen Stiftung and the American Heart Association, Kansas Affiliate KS-94-F-5 to K.-O.N., and by a grant from BioStratum Incorporated to R.G.K. We especially thank Dr. Sripad Gunwar and Ms. Parvin Todd for the help with purification of NC1 and for providing NC1 samples. We also thank Dr. H. Nagase for providing pro-MMP-2, MMP-3 and TIMP-1 for this study and for many helpful discussions. We acknowledge the helpful participation of Ms. Heather Hudson in the initial sequence retrieval and alignment. We wish to thank Dr. W. Bode for communicating the secondary structure assignments of TIMP-1 (see Addendum).

References

- Allan JA, Docherty AJ, Barker PJ, Huskisson NS, Reynolds JJ, Murphy G. 1995. Binding of gelatinases A and B to type-I collagen and other matrix components. *Biochem J* 309:299–306.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410.
- Arnoux B, Merigneau K, Saludjan P, Norris F, Norris K, Bjorn S, Olsen O, Petersen L, Ducruix A. 1995. The 1.6 Å structure of Kunitz-type domain from the alpha 3 chain of human type VI collagen. *J Mol Biol* 246:609–617.
- Barker DF, Pruchno CJ, Jiang X, Atkin CL, Stone EM, Denison JC, Fain PR, Gregory MC. 1996. A mutation causing Alport syndrome with tardive hearing loss is common in the western United States. *Am J Hum Genet* 58:1157–1165.
- Birkendal-Hansen H, Moore WG, Boddien MK, Windsor LJ, Birkendal-Hansen B, DeCarlo A, Engler JA. 1993. Matrix metalloproteinases: A review. *Crit Rev Oral Biol Med* 4:197–250.
- Blumberg B, MacKrell AJ, Fessler JH. 1988. Drosophila basement membrane procollagen alpha 1(IV). II. Complete cDNA sequence, genomic structure, and general implications for supramolecular assemblies. *J Biol Chem* 263:18328–18337.
- Bork P. 1992. Mobile modules and motifs. *Curr Opin Struct Biol* 2:413–421.
- Cecchini JP, Knibiehler B, Mirre C, Le Parco Y. 1987. Evidence for a type-IV-related collagen in *Drosophila melanogaster*. Evolutionary constancy of the carboxyl-terminal noncollagenous domain. *Eur J Biochem* 165:587–593.
- Chu ML, Zhang RZ, Pan TC, Stokes D, Conway D, Kuo HJ, Glanville R, Mayer U, Mann K, Deutzmann R, Timpl R. 1990. Mosaic structure of globular domains in the human type VI collagen alpha 3 chain: Similarity to von

- Willebrand factor, fibronectin, actin, salivary proteins and aprotinin type proteinase inhibitors. *EMBO J* 9:385–393.
- Dion AS, Myers JC. 1987. COOH-terminal propeptides of the major human procollagens. Structural, functional and genetic comparisons. *J Mol Biol* 193:127–143.
- Dölz R, Engel J, Kuhn K. 1988. Folding of collagen IV. *Eur J Biochem* 178:357–366.
- Felsenstein J. 1989. PHYLIP-phylogeny inference package (Version 3.2). *Cladistics* 5:164–166.
- Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol* 266:418–427.
- Gomis-Rüth FX, Maskos K, Betz M, Bergner A, Huber R, Suzuki K, Yoshida N, Nagase H, Brew K, Bourenkov GP, Bartunik H, Bode W. 1997. Mechanism of inhibition of the human matrix metalloproteinase stromelysin-1 by TIMP-1. *Nature* 389:77–81.
- Hobohm U, Sander C. 1995. A sequence property approach to searching protein databases. *J Mol Biol* 251:390–399.
- Hudson BG, Reeders ST, Tryggvason K. 1993a. Type IV collagen: Structure, gene organization, and role in human diseases. Molecular basis of Goodpasture and Alport syndromes and diffuse leiomyomatosis. *J Biol Chem* 268:26033–26036.
- Hudson BG, Kalluri R, Gunwar S, Noelken ME, Mariyama M, Reeders ST. 1993b. Molecular characteristics of the Goodpasture autoantigen. *Kidney Int* 43:135–139.
- Itoh Y, Binner S, Nagase H. 1995. Steps involved in activation of the complex of pro-matrix metalloproteinase 2 (progelatinase A) and tissue inhibitor of metalloproteinases (TIMP)-2 by 4-aminophenylmercuric acetate. *Biochem J* 308:645–651.
- Jones DD. 1975. Amino acid properties and side-chain orientation in proteins: A cross correlation approach. *J Theor Biol* 50:167–183.
- Kahsai TZ, Enders GC, Gunwar S, Brunmark C, Wieslander J, Kalluri R, Zhou J, Noelken ME, Hudson BG. 1997. Seminiferous tubule basement membrane. Composition and organization of type IV collagen chains, and the linkage of alpha3(IV) and alpha5(IV) chains. *J Biol Chem* 272:17023–17032.
- Kalluri R, Gunwar S, Reeders ST, Morrison KC, Mariyama M, Ebner KE, Noelken ME, Hudson BG. 1991. Goodpasture syndrome. Localization of the epitope for the autoantibodies to the carboxyl-terminal region of the alpha 3(IV) chain of basement membrane collagen. *J Biol Chem* 266:24018–24024.
- Kalluri R, Sun MJ, Hudson BG, Neilson EG. 1996. The Goodpasture autoantigen. Structural delineation of two immunologically privileged epitopes on alpha3(IV) chain of type IV collagen. *J Biol Chem* 271:9062–9068.
- Kalluri R, Wilson CB, Weber M, Gunwar S, Chonko AM, Neilson EG, Hudson BG. 1995. Identification of the alpha 3 chain of type IV collagen as the common autoantigen in ant basement membrane disease and Goodpasture syndrome. *J Am Soc Nephrol* 6:1178–1185.
- Kefalides NA, Ohno N, Wilson CB, Fillit H, Zabriski J, Rosenbloom J. 1993. Identification of antigenic epitopes in type IV collagen by use of synthetic peptides. *Kidney Int* 43:94–100.
- Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120.
- Knebelmann B, Breillat C, Forestier L, Arrondel C, Jacassier D, Giatras I, Drouot L, Deschenes G, Grunfeld JP, Broyer M, Gubler MC, Antignac C. 1996. Spectrum of mutations in the COL4A5 collagen gene in X-linked Alport syndrome. *Am J Hum Genet* 59:1221–1232.
- Knight CG, Willenbrock F, Murphy G. 1992. A novel coumarin-labelled peptide for sensitive continuous assays of the matrix-metalloproteinases. *FEBS Lett* 296:263–266.
- Kubota Y, Takahashi S, Nishikawa K, Ooi T. 1981. Homology in protein sequences expressed by correlation coefficients. *J Theor Biol* 91:347–361.
- Leinonen A, Mariyama M, Mochizuki T, Tryggvason K, Reeders ST. 1994. Complete primary structure of the human type IV collagen alpha 4(IV) chain. Comparison with structure and expression of the other alpha (IV) chains. *J Biol Chem* 269:26172–26177.
- Lemmink HH, Schroeder CH, Brunner HG, Nelen MR, Zhou J, Tryggvason K, Haagsma-Schouten WA, Roodvoets AP, Rascher W, van Oost BA, Smeets HJM. 1993. Identification of four novel mutations in the COL4A5 gene of patients with Alport syndrome. *Genomics* 17:485–489.
- Mariyama M, Kalluri R, Hudson BG, Reeders ST. 1992. The alpha 4(IV) chain of basement membrane collagen. Isolation of cDNAs encoding bovine alpha 4(IV) and comparison with other type IV collagens. *J Biol Chem* 267:1253–1258.
- Matsukura H, Butkowski RJ, Fish AJ. 1993. The Goodpasture antigen: Common epitopes in the globular domains of collagen IV. *Nephron* 64:532–539.
- Monboisse JC, Garnotel R, Bellon G, Ohno N, Perreau C, Borel JP, Kefalides NA. 1994. The alpha 3 chain of type IV collagen prevents activation of human polymorphonuclear leukocytes. *J Biol Chem* 269:25475–25482.
- Myers JC, Howard PS, Jelen AM, Dion AS, Macarak EJ. 1987. Duplication of type IV collagen COOH-terminal repeats and species-specific expression of alpha 1(IV) and alpha 2(IV) collagen genes [published erratum appears in *J Biol Chem* 262(29):14372]. *J Biol Chem* 262:9231–9238.
- Nagase H. 1996. Matrix metalloproteinases. In: Hooper, NM, ed. *Zinc metalloproteinases in health and disease*. London, UK: Taylor & Francis. pp. 153–204.
- Nagase H, Fields GB. 1996. Human matrix metalloproteinase specificity studies using collagen sequence-based synthetic peptides. *Biopolymers* 40:339–416.
- Netzer KO, Seibold S, Gross O, Lambrecht R, Weber M. 1996. Use of psoralen-coupled nucleotide primers for screening of COL4A5 mutations in Alport syndrome. *Kidney Int* 50:1363–1367.
- Ninomiya Y, Kagawa M, Iyama K, Naito I, Kishiro Y, Seyer JM, Sugimoto M, Ohashi T, Sado Y. 1995. Differential expression of two basement membrane collagen genes, COL4A6 and COL4A5, demonstrated by immunofluorescence staining using peptide-specific monoclonal antibodies. *J Cell Biol* 130:1219–1229.
- Oberbäumer I, Laurent M, Schwarz U, Sakurai Y, Yamada Y, Vogeli G, Voss T, Siebold B, Glanville RW, Kuhn K. 1985. Amino acid sequence of the non-collagenous globular domain (NC1) of the alpha 1(IV) chain of basement membrane collagen as derived from complementary DNA. *Eur J Biochem* 147:217–224.
- Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85:2444–2448.
- Quinones S, Bernal D, Garcia-Sogo M, Elena SF, Saus J. 1992. Exon/intron structure of the human alpha 3(IV) gene encompassing the Goodpasture antigen (alpha 3(IV)NC1). Identification of a potentially antigenic region at the triple helix/NC1 domain junction. *J Biol Chem* 267:19780–19784.
- Renieri A, Bruttini M, Galli L, Zanelli P, Neri T, Rossetti S, Turco A, Heiskari N, Zhou J, Gusmano R, Massella L, Banfi G, Scolari F, Sessa A, Rizzoni G, Tryggvason K, Pignatti PF, Savi M, Ballabio A, De Marchi M. 1996. X-linked Alport syndrome: An SSCP-based mutation survey over all 51 exons of the COL4A5 gene. *Am J Hum Genet* 58:1192–1204.
- Ries A, Engel J, Lustig A, Kuhn K. 1995. The function of the NC1 domains in type IV collagen. *J Biol Chem* 270:23790–23794.
- Rost B, Sander C. 1993a. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 232:584–599.
- Rost B, Sander C. 1993b. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* 90:7558–7562.
- Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.
- Sander C, Schneider R. 1991. Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins* 9:56–68.
- Saus J, Wieslander J, Langeveld JP, Quinones S, Hudson BG. 1988. Identification of the Goodpasture antigen as the alpha 3(IV) chain of collagen IV. *J Biol Chem* 263:13374–13380.
- Siebold B, Deutzmann R, Kuhn K. 1988. The arrangement of intra- and intermolecular disulfide bonds in the carboxyterminal, non-collagenous aggregation and cross-linking domain of basement-membrane type IV collagen. *Eur J Biochem* 176:617–624.
- Stubbs M, Summers L, Mayr I, Schneider M, Bode W, Huber R, Ries A, Kuhn K. 1990. Crystals of the NC1 domain of human type IV collagen. *J Mol Biol* 211:683–684.
- Timoneda J, Gunwar S, Monfort G, Saus J, Noelken ME, Hudson BG. 1990. Unusual dissociative behavior of the noncollagenous domain (hexamer) of basement membrane collagen during electrophoresis and chromatofocusing. *Conn Tiss Res* 24:169–186.
- Timpl R, Oberbäumer I, von der Mark H, Bode W, Wick G, Weber S, Engel J. 1985. Structure and biology of the globular domain of basement membrane type IV collagen. *Ann NY Acad Sci* 460:58–72.
- Tsilibary EC, Charonis AS, Reger LA, Wohlhueter RM, Furcht LT. 1988. The effect of nonenzymatic glucosylation on binding of the main noncollagenous NC1 domain to type IV collagen. *J Biol Chem* 263:4302–4308.
- Tverskaya S, Bobrynya V, Tsalykova F, Ignatova M, Krasnopolskaya X, Evgrafov O. 1996. Substitution of A1498D in noncollagen domain of alpha 5(IV) collagen chain associated with adult-onset X-linked Alport syndrome. *Hum Mutat* 7:149–150.
- Weber S, Dölz R, Timpl R, Fessler JH, Engel J. 1988. Reductive cleavage and reformation of the interchain and intrachain disulfide bonds in the globular hexameric domain NC1 involved in network assembly of basement membrane collagen (type IV). *Eur J Biochem* 175:229–236.
- Weber S, Engel J, Wiedemann H, Glanville RW, Timpl R. 1984. Subunit structure and assembly of the globular domain of basement-membrane collagen type IV. *Eur J Biochem* 139:401–410.
- Wieslander J, Bygren P, Heinegard D. 1984. Isolation of the specific glomerular basement membrane antigen involved in Goodpasture syndrome. *Proc Natl Acad Sci U.S.A.* 81:1544–1548.

- Wieslander J, Langeveld J, Butkowski R, Jodlowski M, Noelken M, Hudson BG. 1985. Physical and immunochemical studies of the globular domain of type IV collagen. Cryptic properties of the Goodpasture antigen. *J Biol Chem* 260:8564–8570.
- Williamson RA, Martorell G, Carr MD, Murphy G, Docherty AJ, Freedman RB, Feeney J. 1994. Solution structure of the active domain of tissue inhibitor of metalloproteinases-2. A new member of the OB fold protein family. *Biochemistry* 33:11745–11759.
- Zhou J, Barker DF, Hostikka SL, Gregory MC, Atkin CL, Tryggvason K. 1991. Single base mutation in alpha 5(IV) collagen chain gene converting a conserved cysteine to serine in Alport syndrome. *Genomics* 9:10–18.
- Zhou J, Ding M, Zhao Z, Reeders ST. 1994. Complete primary structure of the sixth chain of human basement membrane collagen, alpha 6(IV). Isolation of the cDNAs for alpha 6(IV) and comparison with five other type IV collagen chains. *J Biol Chem* 269:13193–13199.
- Zhou J, Gregory MC, Hertz JM, Barker DF, Atkin C, Spencer ES, Tryggvason K. 1993. Mutations in the codon for a conserved arginine-1563 in the COL4A5 collagen gene in Alport syndrome. *Kidney Int* 43:722–729.
- Zimmerman JM, Eliezer N, Simha R. 1968. The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol* 21:170–201.