Matrix Biology Vol. 14/1994, pp. 233–239

Common Topology within a Non-Collagenous Domain of Several Different Collagen Types

MAHNAZ MORADI-AMÉLI, GILBERT DELÉAGE, CHRISTOPHE GEOURJON and MICHEL VAN DER REST

Institut de Biologie et de Chimie des Protéines, UPR412-CNRS, Lyon, France.

Abstract

The secondary structure of a conserved non-collagenous module in α1(V), α1(XI), α1(IX), α1(XII), α1(XIV) and α1(XVI) collagen chains and in proline- and arginine-rich protein was analyzed using different algorithms. The results predict that a common anti-parallel β-sheet structure composed of nine consensus β-strands is present in these non-collagenous modules. A model for the packing of these β-sheets is proposed which suggests that the predicted β-sheet structure may be involved in molecular recognition functions.

Key words: collagens, protein structure prediction, β-sheet.

Introduction

Collagens are the most abundant fibrous proteins of the animal extracellular matrix. Their common structural motif is the presence of at least one triple-helical domain provided by three α chains containing Gly-X-Y repeats. The fibrillar or interstitial collagens (types I, II, III, V and XI) consist of a central rod-like uninterrupted triple helix and are involved in the formation of striated fibrils. Fibrils containing collagen I as their major constituent also include collagen V and, in soft connective tissues, collagen III. Collagen XI is associated with collagen II in hyaline cartilage fibrils. Other collagen molecules, often called non-fibrillar collagens, have very diverse structures. Among them, the fibril-associated collagens with interrupted triple helices (FACIT)[1] (collagens IX, XII, XIV and XVI) represent a recently described subfamily. It is likely that these latter collagens play a role in the supramolecular organization of extracellular matrix (van der Rest and Garrone, 1991; van der Rest and Bruckner, 1993).

The long central triple helix of fibrillar collagens is flanked by two extensions (N- and C-propeptides), mainly non-collagenous, which are totally or partially processed after their secretion in the extracellular space. FACIT molecules are made of short segments of triple helical structure interspersed by non-collagenous (NC) domains. Despite the structural and functional dissimilarity between FACIT and fibrillar collagens, a striking sequence similarity was previously noted for a module found in the NC domain of FACIT and fibrillar collagens V and XI (Bork, 1992). This module was termed NC4-like domain (Wälchli et al., 1993; Mayne and Brewton, 1993) since it was first described in the NC4 domain of the α1(IX) collagen chain. It is located in the most N-terminal part of the N-propeptide of the α1 chains of collagens V and XI and proximal to the most N-terminal triple helical domain in FACIT collagens. This sequence similarity extends to the proline- and arginine-rich protein (PARP) (Neame et al., 1990) which was recently shown to be part of the N-terminal non-collagenous extension of the pro α2(XI) chain (Zhidkova et al., 1993). The NC4-like domain also has sequence homology with the heparin-binding domain of the adhesive glycoprotein thrombospondin and has therefore been called the tsp1 module by Bork (1992).

The homology noted between these domains suggests a similarity in structure and internal symmetry. In the present work, we used different algorithms to predict the secondary structure of a stretch of 160 amino acid residues within this

[1] Abbreviations used: FACIT, fibril-associated collagens with interrupted triple-helices; NC, non-collagenous; PARP, proline- and arginine-rich protein; Ig, immunoglobulin.

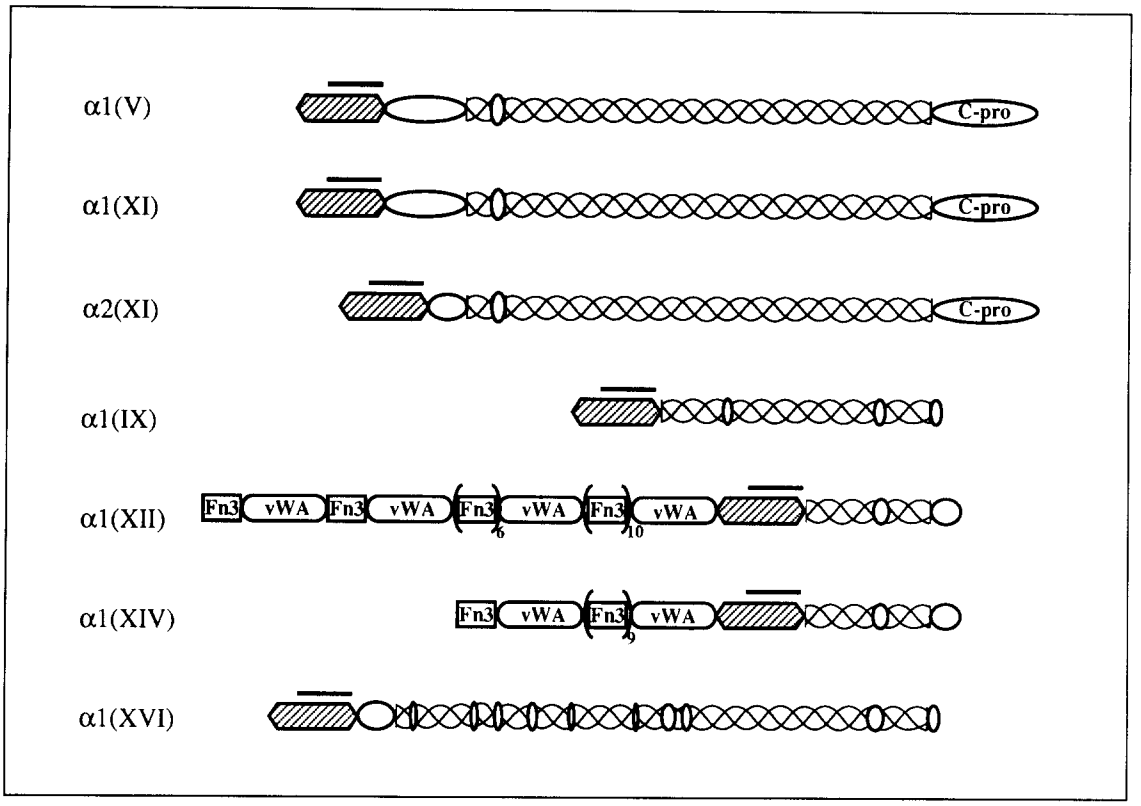

Fig. 1. The position of the NC-module in different collagen types. Collagen molecules are oriented with their N-termini to the left. Hatched boxes represent region of sequence homology (NC4-like domain or tsp1 module) in different collagen types. Thick black lines indicate the position of NC-module within this region. PARP is illustrated as a hatched box in the α2(XI) molecule (Zhidkova et al., 1993). Other non-collagenous domains are represented by open boxes or ovals. Abbreviations used for non-triple-helical regions: C-pro, C-propeptide; Fn3, fibronectin type III repeat; vWA, von Willebrand factor A domain.

module, called hereafter NC-module. A consensus β-stranded pattern could be deduced for this domain in all the mentioned collagen chains. The presence of such a pattern suggests a common function for this module that could potentially be involved in molecular recognition.

Methods

The amino acid sequences of pro-α1(V) collagen (Takahara et al., 1991; Greenspan et al., 1991), pro-α1(XI) collagen (Yoshioka and Ramirez, 1990), α1(IX) collagen (Muragashi et al., 1990) and α1(XVI) collagen (Pan et al., 1992) were deduced from published human cDNA sequences. The α1(XII) (Gordon et al., 1989; Yamagata et al., 1991), α1(XIV) (Trueb and Trueb, 1992; Wälchli et al., 1993) and α1(IX) (Vasios et al., 1988) sequences were from chicken cDNA sequences. The primary structure of PARP was previously published for bovine cartilage (Neame et al., 1990). The location of the NC-module in each collagen type is highlighted in Fig. 1.

All sequence analyses were performed using the ANTHEPROT package on an IBM workstation (Geourjon et al., 1991; Geourjon and Deléage, 1993). The sequences of different collagens within the NC-module were aligned using the CLUSTAL program (Higgins and Sharp, 1988), with values of 10 for both fixed- and variable-gap penalties. From the resulting alignment, a pattern was deduced from all amino acid occurrences for each given position of the multiple alignment. The search for homologous regions in proteins with known structure was carried out using this pattern through windows of variable decreasing lengths (from 15 to 7) by scanning all the sequences extracted from the Brookhaven Data Bank.

The secondary structure of all sequences was predicted using (i) information theory with nil decision constants (Garnier et al., 1978), (ii) independent class prediction (Deléage and Roux, 1987), (iii) sequence similarity as in Levin et al. (1986) with window length of 17 amino acids and similarity threshold value of 7, (iv) rules-based approach (Chou and Fasman, 1978) as implemented by Deléage et al. (1987), (v) self-optimised prediction (Geourjon and Deléage, 1994). The accuracy of predictions of these methods ranges from 56 to 69% for a three-state description of secondary structure (α-helix, β-sheet and aperiodic), but the confidence in a predicted segment is

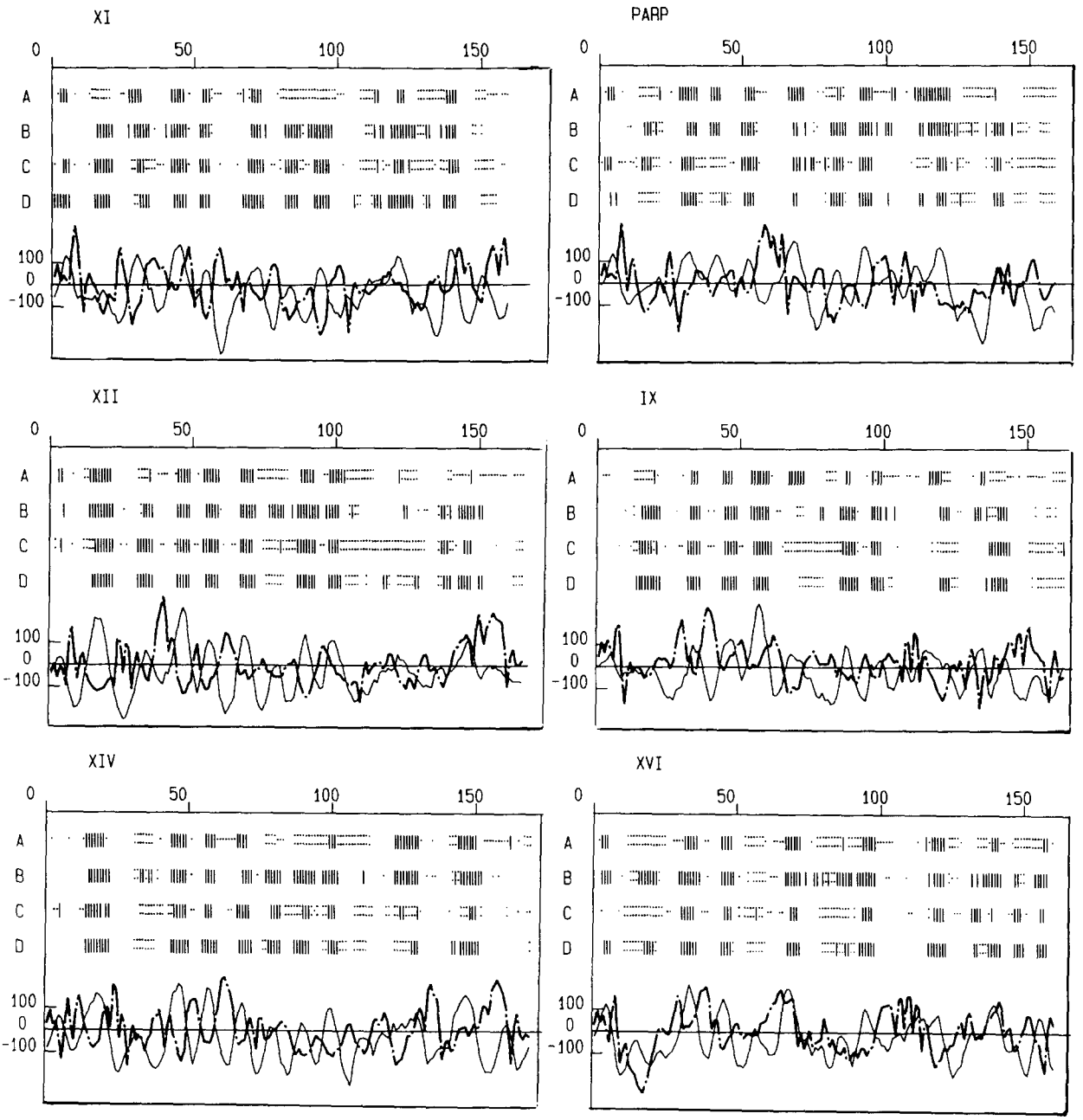

Fig. 2. Predicted secondary structure of conserved NC-module of different collagens. Secondary structures were estimated according to: (A) Garnier et al., 1978; (B) Deléage and Roux, 1987; (C) Levin et al., 1986; (D) Geourjon and Deléage, 1994. Vertical bars (| | | |) denote β-sheet, colons (::::::) indicate α-helix and dots (.....) show β-turn. The profiles shown at the bottom of each panel represent a superimposition of β-sheet (solid lines) and β-turn (bold broken lines) conformational scores according to Garnier et al., 1978. With a 75% sequence identity of the α1(V) with α1(XI) chains, the predicted structure for the two chains is highly similar. Therefore, we only present here the α1(XI) structure. Similarly, human α1(IX) is highly similar to chicken α1(IX) and is not presented.

increased when methods are in agreement. For each protein sequence, the final secondary structure prediction was established on the basis of agreement of three out of five methods. Then, these jointly predicted structures were related to the sequence alignment of homologous regions of the proteins, allowing the emergence of consensus structure motifs.

Results

Different algorithms were used for secondary structure analysis based on primary sequence data of an approximately 160-residue-long NC-module from N-terminal propeptides of α1(V) and α1(XI) (Greenspan et al., 1991; Yoshioka and Ramirez, 1990), of PARP (Neame et al.,

Table I. Comparison between the secondary structures predicted for the collagen NC-module and similar structures of different proteins already elucidated by X-ray crystallography.

Protein	Enolase[a]	Photosynthetic reaction center[b]	Ca^{2+} binding protein[c]	
Sequence	^{242}KIGLDC^{247}	^{42}GDAQIG^{47}	^{17}EGD^{19}	^{57}DGE^{59}
X-ray structure[d]	bbbbbb	TTTbbb	TTT	TTT
α1 chain of	collagen XII	collagen IX	collagen XVI	
Sequence	^{2665}KIYIDC^{2670}	^{138}GKEQVG^{143}	^{150}DGD^{153}	
Predicted structure	bbbbbb	TTTbbb	TTT	

[a] Stec and Lebioda (1990). Brookhaven Data Bank entry code: 2enl.
[b] Deisenhofer and Michel (1989). Brookhaven Data Bank entry code: 1prc.
[c] Szebenyi and Moffat (1986). Brookhaven Data Bank entry code: 3icb.
[d] b and T denote β-sheet and turn structures, respectively.

1990), of NC3 domains of α1(XII) (Gordon et al., 1989) and α1(XIV) (Trueb and Trueb, 1992), of NC4 domain of α1(IX) (Muragaki, 1990; Vasios et al., 1988) and of NC11 domain of α1(XVI) (Pan et al., 1992). Comparison of potential secondary structures showed a good agreement between the different predictive methods. Interestingly, the percentage of β-sheet was always higher than that of α-helix regardless of the predictive method used (Fig. 2). Analysis of the β-sheet and β-turn profiles demonstrated that, at least for the first half of the sequence, the maximal score for β-turn corresponded to a minimal score for β-sheet and *vice versa*, suggesting an alternation of β-sheets and β-turns. In addition, a computer search of consensus sequences on proteins with available X-ray crystallographic data allowed us to confirm the prediction for some of the β-sheets and turns (Table I). It should be kept in mind, however, that in some rare cases identical sequences have been reported not to present identical structures (Kabsh and Sander, 1984).

The joint prediction approach combined with multiple alignment of sequences allowed us to define the presence of β-strand blocks of 5–6 amino acids in length, common to all indicated collagen chains (Fig. 3). Notable conserved patterns include: (i) alternating polar and apolar residues in blocks 1 and 2 with charged residues before the first and after the second block; (ii) a β-strand in block 3 with a conserved Q residue, like the one observed on the beginning of the β-sheet structure reported in photosynthetic reaction

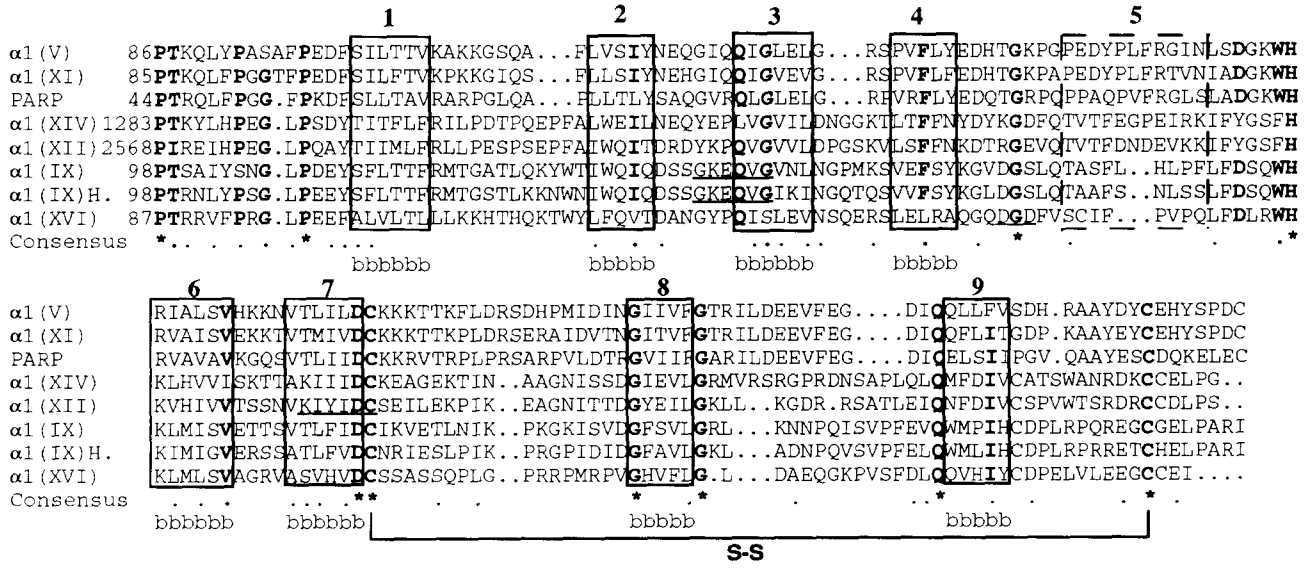

Fig. 3. Sequence/structure alignment of NC-module in N-propeptides of α1(V) and α1(XI), in PARP (α2(XI)), in NC3 domain of α1(XII) and α1(XIV), in NC4 domain of α1(IX) and in NC11 domain of α1(XVI). The nine predicted consensus β-strands (bbbb) are boxed and numbered. In the consensus line, stars indicate identical residues and dots denote conservative changes in amino acids. Identical residues are in bold if they occur in at least six of eight sequences. The known or predicted disulfide bridge is drawn between relevant cysteine residues. The short sequences presented in Table I are underlined.

center (Table I); (iii) a characteristic pattern of two aromatic residues in block 4 (except for collagen XVI); (iv) a conserved G residue in the loop between β-strands in blocks 4 and 5; (v) the presence of a conserved WH sequence before a conserved positively charged residue which begins the β-strand of block 6; (vi) the presence at the end of block 7 of a conserved DC sequence which also ends a β-sheet structure in enolase (Table I) (this C residue has been implicated in disulfide bridge formation in PARP (see below)); (vii) the presence of two conserved G residues in each side of block 8. Among these β-stranded blocks, only the β-structure of block 5 could not be easily demarcated due to the presence of gaps within it.

Some exceptions exist in this overall presentation of β-structures. In fact, block 2 of collagen XIV, block 4 of collagen XVI, block 5 of human collagen IX and block 8 of collagen XII are predicted to present an α-helical structure. Regarding the close sequence homology between α1(XII) (Gordon et al., 1989) and α1(XIV) (Trueb and Trueb, 1992; Wälchli et al., 1993) chains in the region of block 2, the predicted α-helical structure of the latter chain is rather surprising. However, the replacement of E by Q, as observed in the collagen XII sequence, shifted the consensus prediction towards a β-sheet structure (not shown).

In the second half of the sequence, several connecting segments sometimes exhibited short α-helical stretches. Of particular interest are two conserved cysteine residues located adjacent to block 7 and after block 9. These residues are likely to form a disulfide bond, since chemical elucidation of disulfide bonds on PARP demonstrated an intrachain disulfide bridge between these cysteines (Neame et al., 1990).

As previously mentioned, thrombospondin bears some sequence homology with this NC-module (Bork, 1992). The predicted structure of thrombospondin shares a significant structural resemblance with the proposed consensus β-stranded structure (data not shown). However the first two β-strands observed in Fig. 3 are missing from its predicted structure.

Discussion

When a secondary structure analysis is performed, the use of several methods based on different principles allows the precise location of consensus regions. It was shown that the prediction accuracy is increased by making joint predictions (Biou et al., 1988). Using this type of approach, the prediction of secondary structure of the NC-module shared between N-terminal propeptide of α1(V) and α1(XI) collagen chains and the NC domain proximal to the most N-terminal triple-helical domain of FACITs (Fig. 1) was performed. Predictive structural analysis defines nine β-strands in the conserved NC-module of these proteins. These strands are likely to be folded into anti-parallel β-sandwiches as suggested by the alternation of β-sheets and β-turns. The replacement of β-sheet structure by an α-helix in some exceptional cases (4 over 72) may be explained by the fact that the accuracy of secondary structure prediction is not expected to reach 100%. The emergence of β-strands in several homologous domains supports the view that, even in these cases, β-structures may be present.

The presence of seven consensus β-strands was reported in extracellular segment of cytokine receptors by a predictive method (Bazan, 1990). In human fibronectin type III domains, the presence of the predicted seven β-strands was subsequently confirmed by NMR study (Baron et al., 1992). X-ray crystallography of the three-dimensional folding of immunoglobulin (Ig) domains revealed two main structural classes that contain seven to nine β-strands. The

Table II. Comparison of key residues in V-SET and NC-module sequences.

	β-stand D[a]		β-strand F	
V-SET sequences[b]	^{1}R	^{2}[V,I,F,A][c]	$^{-2}$D	^{1}[G,A]
	β-strand 6		β-strand 8	
NC-module sequence[d]	^{1}[R,K]	^{2}[V,I,L]	$^{-1}$[D,N]	^{1}G

[a] The positions of characteristic residues on β-strands D and F are from Williams and Barclay (1988).

[b] V-SET includes, in addition to Ig variable domains, antigen receptor V-domains and other sequences likely to have a V-type fold.

[c] Residues in square brackets indicate amino acids of similar type which occur in the same position in different V-SET or NC-modules. The position relative to each β-strand is numbered.

[d] Most representative sequences of β-strands in NC-modules were taken from Fig. 3.

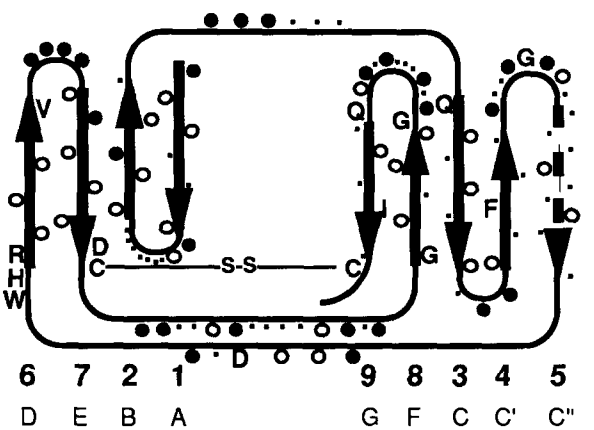

Fig. 4. Hypothetical diagram of the folding patterns of the nine predicted β-strands of the NC-module. Conserved residues are indicated by their one-letter code. Open circles show consensus (> or = 7/8) hydrophobic residues (A, I, L, V, M, W, F). Closed circles indicate consensus hydrophilic residues (D, E, N, Q, T, S, K, R, H). Where no consensus for either polar or hydrophobic residues could be found, residues are indicated by a dot. Under the diagram, the equivalent β-strands in an Ig V_H subunit are marked as indicated by Williams and Barclay (1988).

nine-stranded Ig fold of variable (V_H and V_L) subunit consisted of two twisted, stacked β-sheets composed of four (A, B, E, D) and five (C, C′, C″, F, G) strands linked by a disulfide bridge between strands B and F (Amzel and Poljak, 1979; Williams and Barclay, 1988). The conserved NC-module presents some resemblance to the V region of Ig. As the NC-module in all collagens except XII and XIV, this region occurs at the N-terminal end of the molecule, is approximately 110 amino acid residues long and is composed of a nine-stranded β-sheet. Conserved hydrophilic residues at the beginning or at the end of strands 1, 2 and 6 have their counterpart in the sequence templates postulated for β-strands A, B and D of variable Ig domain (Bazan, 1990). In regions outside or in the beginning of β-strands D and F, some conserved patterns characteristic of Ig V domain and called V-SET sequences were reported by Williams and Barclay (1988). A likely similar pattern is present in the conserved NC-module, as indicated in Table II.

Based on these observations, we used the V_H region as a model to propose in Fig. 4 the disposition of nine β-strands which can be folded into two β-sheets in the conserved NC-module of collagens. The two anti-parallel β-sheets would be composed of four (1, 2, 6, 7) and five (3, 4, 5, 8, 9) β-strands, respectively. Taking the spatial arrangement into consideration, the disulfide bridge between the highly conserved cysteine residue of block 7 and that found after block 9 can be considered as an equivalent of the disulfide bond observed in the Ig-fold domain (between strands B and F) that keeps the paired β-sheets close together. The conserved phenylalanine in β-strand 4 can be considered as a spatial equivalent of the characteristic tryptophan residue on Ig strand C. This model is supported by the presence of hydrophobic residues within the interior of the anti-parallel strands and by the fact that the loops are mostly composed of hydrophilic residues (Fig. 4). However, it should be noted that the validity of this proposed model can only be confirmed by crystallography or NMR studies.

Other β-stranded structures have been observed in proteins that are involved in molecular recognition and exert adhesion or binding functions. The folding of the functional domain of Ig-related molecules can be considered as providing a stable platform for the presentation of specific determinants involved in molecular recognition mechanisms. The topological resemblance of the proposed three-dimensional folding of NC-module to the functional domain of Ig potentially suggests a similar interactive function for this module. Such a function has been postulated for the NC4 domain of the α1(IX) collagen chain which is thought to interact with glycosaminoglycan (Shaw and Olsen, 1991).

Acknowledgement

The authors wish to thank Professor Richard Mayne for critical reading of the manuscript.

References

Amzel, L.M. and Poljak, R.J.: Three-dimensional structure of immunoglobulins. *Ann. Rev. Biochem. 48:* 961–997, 1979.

Baron, M., Main, A.L., Driscoll, P.C., Mardon, H.J., Boyd, J. and Campbell, I.D.: ^{1}H NMR assignment and secondary structure of the cell adhesion type III module of fibronectin. *Biochem. 31:* 2068–2073, 1992.

Bazan, J.F.: Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Natl. Acad. Sci. USA 87:* 6934–6938, 1990.

Biou, V., Gibrat, J.F., Levin, J.M., Robson, B. and Garnier, J.: Secondary structure prediction: combination of three different methods. *Prot. Eng. 2:* 185–191, 1988.

Bork, P.: The molecular architecture of vertebrate collagens. *FEBS Lett. 307:* 49–54, 1992.

Chou, P.Y. and Fasman, G.D.: Prediction of secondary structure of proteins from amino acid sequence. *Adv. Enzymol. Relat. Subj. Biochem. 47:* 45–148, 1978.

Deisenhofer, J. and Michel, H.: The photosynthetic reaction center from the purple bacterium *Rhodopseudomonas viridis*. *Science 245:* 1463–1473, 1989.

Deléage, G. and Roux, B.: An algorithm for protein secondary structure prediction based on class prediction. *Prot. Eng. 1:* 289–294, 1987.

Deléage, G., Tinland, B. and Roux, B.: A computerized version of the CHOU and FASMAN method for predicting the secondary structure of proteins. *Anal. Biochem. 163:* 292–297, 1987.

Garnier, J., Osguthorpe, D.J. and Robson, B.: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol. 120:* 97–120, 1978.

Geourjon, C., Deléage, G. and Roux, B.: ANTHEPROT: An interactive graphic software for analyzing protein structure. *J. Mol. Graphics 9:* 188–190, 1991.

Geourjon, C. and Deléage, G.: Interactive and graphic coupling between multiple alignments, secondary structures prediction and motif/pattern scanning into protein sequences. *Comput. Appl. Biosci. 9:* 87–91, 1993.

Geourjon, C. and Deléage, G.: SOPM: a self-optimized method for predicting the secondary structure. *Prot. Eng.7:* 157–164, 1994.

Gordon, M.K., Gerecke, D.R., Dublet, B., van der Rest, M. and Olsen, B.R.: Type XII collagen. A large multidomain molecule with partial homology to type IX collagen. *J. Biol. Chem. 264:* 19772–19778, 1989.

Greenspan, D.S., Cheng, W. and Hoffmann, G.G.: The pro-α1(V) collagen chain. Complete primary structure, distribution of expression, and comparison with the pro-α1(XI) collagen chain. *J. Biol. Chem. 266:* 24727–24733, 1991.

Higgins, D.G. and Sharp, P.M.: Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene 73:* 237–244, 1988.

Kabsh, W. and Sander, C.: On the use of sequence homologies to predict protein structure: Identical pentapeptide can have completly different conformations. *Proc. Natl. Acad. Sci. USA 81:* 1075–1078, 1984.

Levin, J.M., Robson, B. and Garnier, J.: An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett. 205:* 303–308, 1986.

Mayne, R. and Brewton, R.G.: New members of the collagen superfamily. *Curr. Opin. Cell Biol. 5:* 883–890, 1993.

Muragaki, Y., Kimura, T., Ninomiya, Y. and Olsen, B.R.: The

complete primary structure of two distinct forms of human α1(IX) collagen chains. *Eur. J. Biochem. 192:* 703–708, 1990.

Neame, P. J., Young, C. N. and Treep, J. T.: Isolation and primary structure of PARP, a 24-kDa proline- and arginine-rich protein from bovine cartilage closely related to the NH_2-terminal domain in collagen α1(XI). *J. Biol. Chem. 265:* 20401–20408, 1990.

Pan, T. C., Zhang, R. Z., Mattei, M. G., Timpl, R. and Chu, M. L.: Cloning and chromosomal location of human α1 (XVI) collagen. *Proc. Natl. Acad. Sci. USA 89:* 6565–6569, 1992.

Shaw, L. M. and Olsen, B. R.: FACIT collagens: diverse molecular bridges in extracellular matrix. *Trends Biochem. Sci. 16:* 191–194, 1991.

Stec, B. and Lebioda, L.: Refined structure of yeast Apo-enolase at 2.25 Å resolution. *J. Mol. Biol. 211:* 235–248, 1990.

Szebenyi, D. M. E. and Moffat, K.: The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine. Molecular details, ion binding, and implications for the structure of other calcium-binding proteins. *J. Biol. Chem. 261:* 8761–8777, 1986.

Takahara, K., Sato, Y., Okazawa, K., Okamoto, N., Noda, A., Yaoi, Y. and Kato, I.: Complete primary structur* of human collagen α1(V) chain. *J. Biol. Chem. 266:* 13124–13129, 1991.

Trueb, J. and Trueb, B.: Type XIV collagen is a variant of undulin. *Eur. J. Biochem. 207:* 549–557, 1992.

van der Rest, M. and Garrone, R.: Collagen family of proteins. *FASEB J. 5:* 2814–2823, 1991.

van der Rest, M. and Bruckner, P.: Collagens: diversity at the molecular and supramolecular levels. *Curr. Opin. Struct. Biol. 3:* 430–436, 1993.

Vasios, G., Nishimura, I., Konomi, H., van der Rest, M., Ninomiya, Y. and Olsen, B. R.: Cartilage type IX collagen-proteoglycan contains a large amino-terminal globular domain encoded by multiple exons. *J. Biol. Chem. 263:* 2324–2329, 1988.

Wälchli, C., Trueb, J., Kessler, B., Winterhalter, K. H. and Trueb, B.: Complete primary sequence of chicken collagen XIV. *Eur. J. Biochem. 212:* 483–490, 1993.

Williams, A. F. and Barclay, A. N.: The immunoglobulin superfamily. Domains for cell surface recognition. *Ann. Rev. Immunol. 6:* 381–405, 1988.

Yamagata, M., Yamada, K. M., Yamada, S. S., Shinomura, T., Tanaka, H., Nishida, Y., Obara, M. and Kimata, K.: The complete primary structure of type XII collagen shows a chimeric molecule with reiterated fibronectin type III motifs, von Willebrand factor A motifs, a domain homologous to a noncollagenous domains with an Arg-Gly-Asp site. *J. Cell Biol. 115:* 209–221, 1991.

Yoshioka, H. and Ramirez, F.: Pro-α1(XI) collagen. Structure of the amino-terminal propeptide and expression of the gene in tumor cell lines. *J. Biol. Chem. 265:* 6423–6426, 1990.

Zhidkova, N. I., Brewton, R. G. and Mayne, R.: Molecular cloning of PARP (Proline/arginine-rich protein) from human cartilage and subsequent demonstration that PARP is a fragment of the NH_2-terminal domain of the collagen α2(XI) chain. *FEBS Lett. 326:* 25–28, 1993.

Dr. Mahnaz Moradi-Améli, Institut de Biologie et de Chimie des Proteines, UPR 412-CNRS, 7 passage du Vercors F-69367, Lyon cedex 07, France.

Received October 19, 1993; accepted December 7, 1993.