Volume 303, number 2,3, 126–128 FEBS 11083 June 1992

The fibrillar collagens, collagen VIII, collagen X and the C1q complement proteins share a similar domain in their C-terminal non-collagenous regions

Andy Brass, Karl E. Kadler, J. Terrig Thomas, Michael E. Grant and Raymond P. Boot-Handford

Department of Biochemistry and Molecular Biology, University of Manchester, Oxford Road, Manchester M13 9PT, UK

Received 9 April 1992

A sequence comparison of the C-termini of collagens X, VIII, the collagen-like complement factor C1q, and the fibrillar collagens showed a conserved cluster of aromatic residues. This conserved cluster was in a domain of approximately 130 amino acids that exhibited marked similarities in hydrophilicity profiles between the different collagens, despite a low level of sequence similarity. These data suggest that the 'collagen X-like family' and the fibrillar collagens contain a domain within their C-termini that adopts a common tertiary structure, and that a conserved cluster of aromatic residues in this domain may be involved in C-terminal trimerization.

Collagen; Collagen biosynthesis; Collagen trimerization; Protein folding

1. INTRODUCTION

All collagens and collagen-like molecules such as C1q, lung surfactant protein, acetylcholinesterase and mannose binding protein share the common structural motif of a triple helix (see [1] for recent review). During assembly of collagens, precise alignment of the three chains occurs at the C-terminal end of the helix [1] and is maintained as folding proceeds in a C- to N-direction [2]. Collagens are synthesized as procollagens in which the triple helical domain is flanked by C- and N-terminal globular extensions. A variety of approaches have led to the view that the precise alignment of the three chains is controlled in most, if not all cases by the globular C-terminus [3–6]. Although it is now generally accepted that association of three chains at their non-collagenous C-termini is the prerequisite for correct alignment, nucleation and folding of the collagen triple helix, the mechanism by which this very precise association takes place remains obscure.

2. MATERIALS AND METHODS

Sequence searches and manipulation were carried out using the SEQNET facilities at Daresbury Laboratory, UK. Protein sequences were taken from the OWL non-redundant protein sequence database [7]. Homology scans were performed using the SWEEP program [7]. Pair-wise sequence alignments were performed using the CLUSTAL [8] program with gap penalty parameters set to minimise the number of gaps appearing in the aligned sequences.

Aligned collagen sequences were combined to generate consensus matrices for the conserved aromatic motif (see Results). These matrices were then used to scan the OWL database using the LUPES package [7]. Hydrophilicities of the protein sequences were calculated as described by Kyte and Doolittle [9] and smoothed using a window-size of seven residues. To compare the hydrophilicity profile of the C-terminus of collagen X with profiles of other proteins, plots were aligned based on the sequence comparisons. Where gaps or deletions were introduced into either sequence during alignment, the longest stretch of homology lacking such deletions or gaps was used to align the hydrophilicity plots.

To determine how closely the aligned hydrophilicity profile of the C-terminus of collagen X matched that of other proteins, the root-mean-square (RMS) distance between the two curves was calculated. Statistical comparisons were conducted by both parametric (Student's *t*-test) and non-parametric (Mann-Whitney *U*-test) methods. The *P* values quoted apply to both methods of analysis.

Correspondence address: A. Brass, Department of Biochemistry and Molecular Biology, University of Manchester, Oxford Road, Manchester M13 9PT, UK. Fax: (44) (61) 275-5082.

3. RESULTS AND DISCUSSION

Six proteins showed significant sequence homology with the C-terminal domain of human collagen α1(X): the bovine and chick collagen α1(X) chains; the rabbit collagen α1(VIII) chain, the human collagen α2(VIII) and the human C1qA, C1qB and C1qC complement chains. The area of homology extended over approximately 130 residues (residues 547–680 of collagen α1(X) [12]; residues 116–249 of the B chain of human C1q [13], one region of which showed particularly high homology. Table I shows this region of best alignment. The alignment was then used to create a consensus matrix (not shown) with which to search the OWL database for related sequences. A match was identified between parts of this consensus sequence and sequences found in the C-terminal non-collagenous domain of the fibrillar collagens [collagen α1(I), α2(I), α1(II), α1(III) and α2(V)].

Published by Elsevier Science Publishers B.V.

Table I

```
HVIII(α2)    VKFDRTLYNGHSGYNPATGIFTCPVGGVYYFAYHV
RVIII(α1)    IKFDRLLYNGRQNYNPQTGIFTCEVPGVYYFAYHV
CX    (α1)   IKFDKILYNRQQHYDPRTGIFTCRIPGLYYFSYHA
BX    (α1)   IPFDKILYNKQQHYDPRTGIFTCKIPGIYYFSYHI
HX    (α1)   IPFDKILYNRQQHYDPRTGIFTCQIPGIYYFSYHV
H C1qB       IRFDHVITNMNNNYEPRSGKFTCKVPGLYYFTYHV
H C1qA       I-FDTVITNQEEPYQNHSGRFVCTVPGYYYFTFQV
H C1qC       IRFNAVLTNPQGDYDTSTGKFTCKVPGLYYFVYHA
             + ↓·  + ‖ ·  ↓·  ·‖ ↓ ¦ + ‖+↓↓↓ _·+
```

Alignment of most homologous regions from the C-terminal non-collagenous domains of human α2(VIII) collagen (HVIII), rabbit α1(VIII) collagen (RVIII), α1(X) collagen from chick (CX), bovine (BX) and human (HX) and the B, A and C chains of human C1q. The positions of identical residues are indicated in outline type and by (¦); conserved hydrophobic residues (+); conserved hydrophilic/H-bonding residues (·); conserved aromatic residues (−).

As an example of the matches obtained, the region of homology identified between the human α1(X) and human α1(II) collagens is shown below:

```
Human α1(X)     FTCQIPGLYYFSY
                |   | | | |||        (1)
Human α1(II)    FGETINGGFHFSY
```

We were interested to note that the majority of conserved amino acids in this region were aromatic. No other significant sequence homologies were observed between the C-termini of α1(X) and any of the fibrillar collagens. The new set of alignments was used to refine the consensus matrix and rescan the data base. The top 22 matches with the refined consensus matrix consisted exclusively of the chains of collagen X, collagen VIII, C1q and the fibrillar collagens from various species (data not shown). The consensus sequence for the motif is shown below where x represents any amino acid, and positions at which a choice of amino acids is allowed are shown with the permitted residues in brackets:

Fxxx(VLIM)xGx(FY)xFxY (2)

The hydrophilicity profiles for the C-terminal domains of collagens α1(X) and α1(II) were calculated with no gaps or deletions and the plots aligned based on the sequence alignment shown (see alignment (1) above). The profiles appear to be well-matched over a segment of approximately 130 amino acid residues flanking the aromatic motif, despite the lack of any obvious sequence homology between the α1(X) and α1(II) chains in the flanking regions (the alignment used to create Fig. 1 starts from residues 29 and 91 of the C-terminal non-collagenous domains of α1(X) and α1(II) collagens, respectively). The region over which the hydrophilicity plots matched was in the same region of the collagen X C-terminus as the 130 amino acid region which was found to be conserved between members of the collagen α1(X)-like family described earlier.

In order to determine the significance to attach to this unexpected match in the hydrophilicity profiles over a relatively long stretch of the C-terminal non-collagenous domains of collagens X and II (Fig. 1), the probability of obtaining such a match was calculated. Twenty-five different proteins were picked at random from the Brookhaven database of crystallographic structures. Each of these were then aligned, based on sequence, with the collagen X C-terminal domain using the same method described above and the hydrophilicity curves calculated. The RMS distance between the aligned hydrophilicity curves was then determined. The mean of the RMS scores was 1.48 with a standard deviation (SD) of 0.16. Similar calculations were performed comparing collagen X to the homologous C-terminal domains [collagens α1(VIII), α2(VIII), C1qA, C1qB and C1qC] (average RMS = 0.95, SD = 0.18), against the fibrillar collagens [collagens α1(I), α2(I), α1(II), α1(III) and α2(V)] (average RMS = 1.13, SD = 0.08), and against the non-fibrillar collagens [collagens IV, VI, IX and XII] (average RMS = 1.37, SD = 0.06).

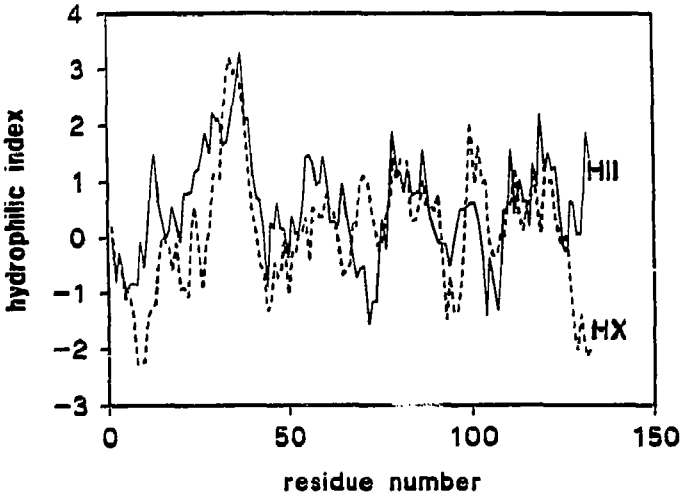

Fig. 1. Alignment of the hydrophilicity profiles for the C-terminal domains of human collagens α1(X) (labelled HX on the figure) and α1(II) (labelled HII on the figure).

The average RMS score for the hydrophilicity plot of collagen X compared to other members of the collagen X-like family was significantly lower than that obtained for the randomly selected proteins or the non-fibrillar collagens ($P<0.005$). The average RMS score for collagen X compared to the fibrillar collagen distribution was also significantly lower than the average RMS score obtained aligning collagen X with randomly selected proteins or non-fibrillar collagens ($P<0.005$). This demonstrates that the aligned hydrophilicity curves for the C-termini of the α1 chains of collagens X and II presented in Fig. 1 (and indeed, between collagen X and the fibrillar collagens in general) are a significantly better match than would be expected at random, whereas the match against the non-fibrillar collagens was not significantly different from random.

These results strongly suggest that the fibrillar and 'collagen X-like' collagens share a structurally conserved C-terminal domain of approximately 130 amino acids. Although the sequence similarity between the two domains is negligible, the match in the hydrophilicity profiles is highly significant. Such 'cryptic' similarities are not unusual for there are many examples of proteins sharing a common folding pattern even though they have almost no residues in common (a case in point being the immunoglobulin superfamily).

If the C-terminal domains in the fibrillar and 'collagen X-like' collagens do indeed share a common structure, it is interesting to speculate on the role of the common domain, and in particularly on the importance of the conserved aromatic motif (2), the only region of sequence similarity conserved between (and unique to) these two families of collagens.

Whilst much has been learnt about the folding of collagen [1], little is known about the initial stages of the trimerization mechanism occurring at the C-terminus. Given that each of these molecules assemble to form a collagenous triple helix, it is reasonable to hypothesize that the most conserved region in their C-termini would be the surface with which each chain physically interacts with its two partners during the first stage of trimerization. The conserved motif (2) within this domain contains four conserved aromatic residues. Hydrophobic interactions are known to be a driving force for protein folding and association, as is seen in the assembly of collagen fibrils [14]. One of the most favoured types of hydrophobic interaction is that between aromatic amino acids as has been demonstrated to occur in the recognition of antigens by antibodies [15] and also in the interaction surfaces of another trimeric protein, tumour necrosis factor, in which 8 out of 16 aromatic residues are found at the trimer interface [16]. We can therefore hypothesize that the conserved 130 amino acid domain identified above is involved in the initial C-terminal trimerisation of collagens, and that the dominant interactions driving the trimerisation involve a small cluster of highly conserved aromatic amino acids.

The hypothesis that collagen trimerization may involve the conserved aromatic residues within a similarly folded trimerization domain of approx. 130 amino acids can, at present, be applied to a number of collagenous molecules (collagens I, II, III, V, VIII, X, XI and C1q). Other collagen types such as collagens IV, VI, IX and XII do not fit in with the hypothesis. Using site-directed mutagenesis it should be possible to test whether or not the conserved aromatic residues identified above play an important role in the trimerisation of some collagen types.

REFERENCES

[1] Engel, J. and Prockop, D.J. (1991) Annu Rev. Biophys. Chem. 20, 137–152.
[2] Bonadio, J.F. and Beyers, P.H. (1985) Nature 316, 363–366.
[3] Traub, W. and Piez, K.A. (1971) Adv. Protein Chem. 25, 243–352.
[4] Uitto, J. and Prockop, D.J. (1974) Biochemistry 13, 4586–4591.
[5] Schofield, J.D., Uitto, J. and Prockop, D.J. (1974) Biochemistry 13, 1801–1806.
[6] Rosenbloom, J., Endo, R. and Harsch, M. (1976) J. Biol. Chem. 251, 2070–2076.
[7] Akrigg, D., Bleasby, A.J., Dix, N.I.M., Findlay, J.B.C., North, A.C.T., Parry-Smith, D.J., Wootton, J.C., Blundel, T.C., Gardner, S.P., Hayes, F., Islam, S., Sternberg, M.J.E., Thornton, J.M., Tickle, I.J. and Murray-Rust, P.A. (1988) Nature 335, 745–746.
[8] Higgins, D.G., Bleasby, A.J. and Fuchs, R. (1992) (in press).
[9] Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105–132.
[10] Dion, A.S. and Myers, J.C. (1987) J. Mol. Biol. 193, 127–143.
[11] Thomas, J.T., Kwan, A.P.L., Grant, M.E. and Boot-Handford, R.P. (1991) Biochem. J. 273, 141–148.
[12] Thomas, J.T., Cresswell, C.J., Rash, B., Nicholi, H., Jones, T., Solomon, E., Grant, M.E. and Boot-Handford, R.P. (1991) Biochem. J. (in press).
[13] Sellar, G.C., Blake, D.J. and Reid, K.B.M. (1991) Biochem. J. 274, 481–490.
[14] Kadler, K.E., Hojima, Y. and Prockop, D.J. (1987) J. Biol. Chem. 262, 15696–15701.
[15] Padlam, E.A. (1990) Proteins: Structure, Function and Genetics 7, 112–124.
[16] Eck, M.J. and Sprang, S.R. (1989) J. Biol. Chem. 264, 17595–17605.