Volume 307, number 1, 49–54 FEBS 11271 July 1992


*Minireview*

# The modular architecture of vertebrate collagens

Peer Bork

*EMBL, Meyerhofstr. 1, 6900 Heidelberg, and Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Str. 10, 1115 Berlin, Germany*

Received 27 April 1992; revised version received 26 May 1992

Collagens are typical mosaic proteins containing a number of shuffled domains. These domains have been classified by sequence similarity in order to characterize their structural and functional relationships to other proteins. This analysis provides an overview of homologies of collagen domains. It also reveals two new relationships: (i) a module common to type V, IX, XI, and XII collagens was found to be homologous to the heparin binding domain of thrombospondin; (ii) the modular architecture of a human type VII collagen fragment was identified. Its N-terminal globular domain contains fibronectin type III repeats located adjacent to a Von Willebrand factor type A module. The proposed structural similarities point to analogous subfunctions of the respective domains in otherwise distinct proteins.

Homology; Mosaic protein; Type VII collagen; Thrombospondin

## 1. INTRODUCTION

Collagens are a group of structural proteins of the extracellular matrix characterized by various Gly-X-Y-repeats which form large triple helices [1,2]. Each collagen consists of three (identical or different) interacting chains. The triple helix-forming parts are surrounded by noncollagenous (NC) domains. To date, the presence of at least 14 different vertebrate collagens (numbered in succession of discovery) has been proved and the existence of several others can be assumed [1]. In addition, an increasing number of invertebrate collagens has been sequenced, only some of them being homologous in their NC domains to those of vertebrates.

Classifications of collagens are mainly based on their supramolecular structure which in turn mirrors the constitution of the helical parts [3]. They can be treated as molecular rods that physically separate the NC domains. Even if the triple helical parts represent the most striking feature of collagens (see definition), tissue specificity as well as defined binding of noncollagens seem to be encoded in the NC domains [4].

These globular NC domains have features of mobile modules, [5] which are widespread in proteins of diverse function. The triple helical segments can also be considered as independent modules since they are found in noncollagens (Table I, [6,7]). Therefore, collagens represent typical mosaic proteins containing a number of domains fused together by exon shuffling [5,8–11]. With the increasing amount of primary structures stored in public databases, the number of puzzling relationships between common modules in otherwise distinct proteins grows rapidly. Thus, various homologies have been reported between collagen domains and modules of noncollagens (e.g. see refs. in Table I and Fig. 1), often simultaneously by different groups. In order to obtain a comprehensive and automatically derived overview of the modular architecture of collagens, to classify the modules according to their sequence similarities, and thus to get some hints about common features, a systematic protein sequence analysis of all available vertebrate collagen sequences was performed.

*Abbreviations:* NC, noncollagenous; Fn3, fibronectin type III; VWA, von Willebrand factor type A; PARP, proline- and arginine-rich protein.

*Correspondence address:* P. Bork, EMBL, Meyerhofstr. 1, 6900 Heidelberg, Germany.

## 2. AUTOMATICALLY DERIVED OVERVIEW OF THE DOMAIN ASSEMBLY IN COLLAGENS

Sequence analysis of all available collagens indicates a complex arrangement of domains (Fig. 1), which are related to numerous adhesive proteins of diverse function (Table I). The detected sequence similarities are consistent with numerous homologies reported over the last ten years (see e.g. refs. in Fig. 1a and Table I). For example, close overall homologies between chains of different collagen types have been observed among the fibrillar collagens (for review see [1,3]). The α1(V) and α1(XI) chains are surprisingly similar to each other and differ from the other fibrillar collagens only in the N-terminal NC domains [12]. Types IX and XII (fibril-

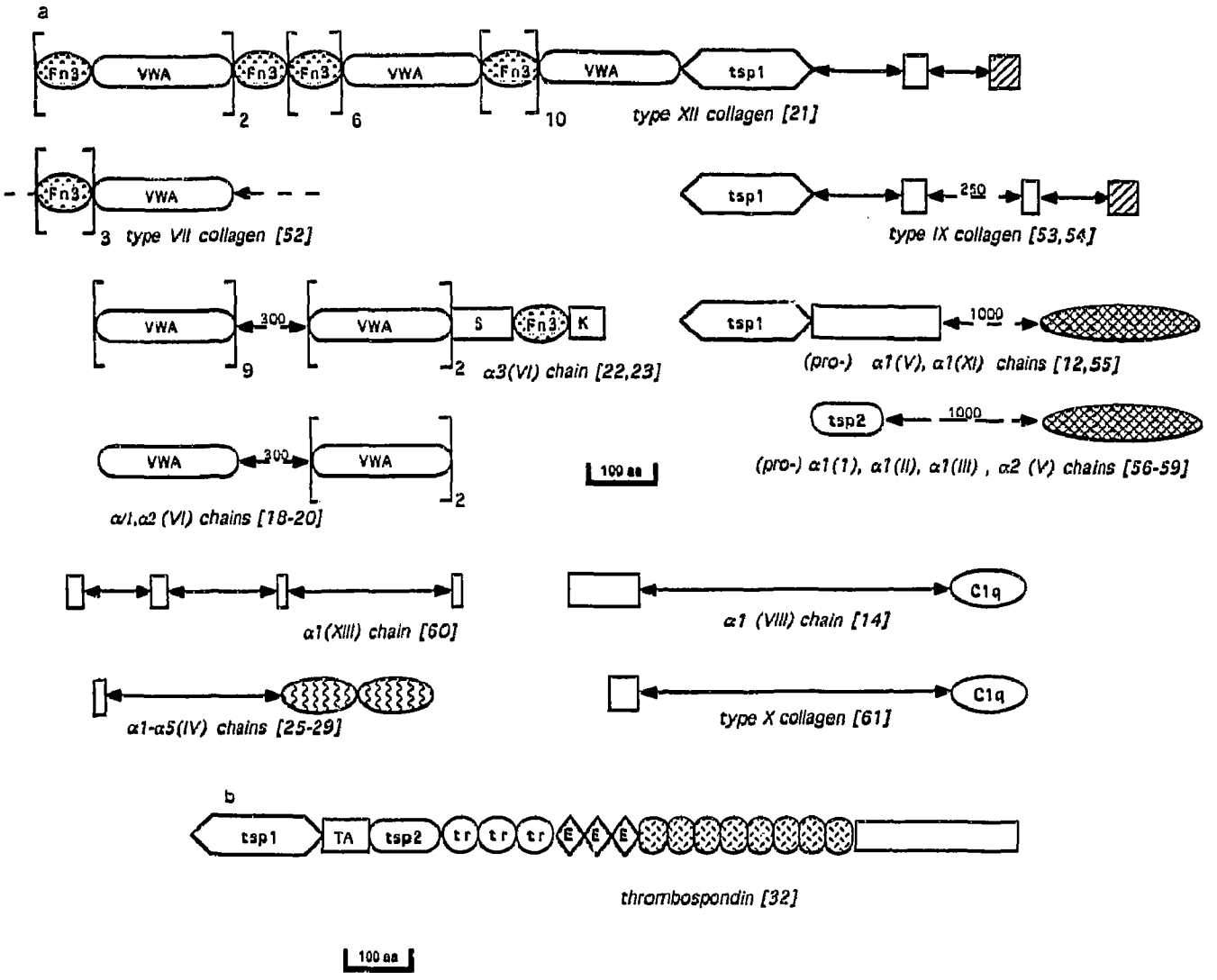


Fig. 1. (a) Overview of noncollagenous modules in collagens. The boxes represent the result of a systematic sequence homology search: All collagen sequences were extracted from SWISS-PROT (release 20; [62]) except for types XII and VII which were directly taken from the literature [21,52]. Initial sequence database searches with the NC domain were carried out using FASTA [48]. Significant hits were aligned by the program PILEUP of the GCG package [63] and the corresponding modules were mapped to the respective collagens. If several sequence segments corresponded to a certain module, motifs were defined and subjected to property pattern searches [50]. In order to improve the sensitivity of the property patterns, the procedure was conducted iteratively, putative homologues being added to the learning set (the multiple alignment) progressively [64]. The significance of the homologies was assessed by using three different homology search methods (FASTA [48], PROFILESEARCH [51], PAT [50]) as described in [65]. The abbreviations of the modules are explained in Table I together with the respective similarities to noncollagenous proteins. Hatched domains have not yet been identified in noncollagens. The white boxes were not found to be similar to any other protein segment. The arrows indicate the triple helical regions (where their full length is not shown the number of amino acids is given). Type VII collagen is only a fragment [52]. Not shown are type XIV collagen (small segments are already sequenced indicating a close relationship to types XII and IX [66]) and the α2(I) chain (similar to collagen the α1(I) chain but lacking the N-terminal thsp2 domain). (b) Modular architecture of thrombospondin. The repeats of thrombospondin are similar to malaria proteins (tr) and epidermal growth factor (E). Type III repeats (hatched) contain calcium binding sites [32] and the TA-region contains two cysteines forming interchain disulfide bridges that are essential for trimer assembly.

associated collagens) were grouped together because they share interrupted triple helices and a globular domain in equivalent locations with respect to their collagenous segments [13]. Also, the α1(VIII) and α1(X) chains have an overall similarity except for their N-terminal regions [14].

Most of the domains in collagens can be aligned with modules of noncollagenous mosaic proteins, e.g. thrombospondin [15], complement component C1q [16,17], Von Willebrand factor [18–21], fibronectin [21–23], Kunitz type inhibitor [22], and several proteins containing collagenous segments (see Table I). Only some domains are exclusively found in collagens such as the duplicated module [24] at the C-terminus of type IV collagen [25–30] or the large NC1 domain common to all fibrillar collagens which are important for chain as-

Table I

Relationships of NC domains in collagens to other proteins

| Symbols | Noncollagens containing corresponding modules |
|---|---|
| Thsp1 | – Thrombospondin (N-terminal heparin binding domain), PARP [31]. |
| Thsp2 | – Thrombospondin (between N-terminus and type I repeat), von Willebrand factor type C module [21]. |
| C1q | – Complement component C1q (also located next to a collagenous segment) [16,17]. |
| Fn3 | – Fibronectin type III module, present in more than 60 different proteins not counting species redundancies. Subfamilies include adhesive matrix proteins such as fibronectin, tenascin, and undulin, receptor protein kinases as well as phosphatases, cytokine receptors, neural adhesive proteins, giant muscle proteins, bacterial carbohydrate-splitting enzymes, and further unclassified proteins like the one causing Kallmann's syndrome (for reviews see [5,10,11]). |
| VWA | – Von Willebrand factor type A module, also found in cartilage matrix protein, several integrin α chains (so-called I-domains), undulin, the regulatory chains of inter-α-trypsin inhibitor, thrombospondin related malaria protein and probably DHP-sensitive voltage dependent Ca-channels [40,41]. |
| K | – Kunitz type inhibitor, also found in amyloid precursor protein, inter-α-trypsin inhibitor [22]. |
| S | – Module found in several salvage proteins [22]. |
| Gxy | – Collagenous repeats also present in scavenger receptor [44], a virus protein [7], asymmetrical acetylcholinesterase, surfactant proteins, several proteins which also contain a lectin type C domain ([44–46] and refs. therein), complement component C1q [16,17] and bacterial pullanase [47]. |

sociation (for review see [3]; Fig. 1). Both are more conserved than domains with similarities to noncollagens (data not shown). NC domains that are homologous among distinct collagen chains and types always have a similar location with respect to the triple helices (Fig. 1). This points to equivalent functions in the different collagen types. Except for types IV, VIII, X and XIII, all collagens (including fibrillar and nonfibrillar ones) can be grouped together by sharing certain globular modules (Fig. 1). Since those domains are thought to fine-tune both binding and tissue specificities, new detected homologies, as described below, may provide some further insight into the functional network of matrix proteins.

## 3. A DOMAIN COMMON TO TYPES V, IX, XI, XII COLLAGENS, AND THROMBOSPONDIN

A large domain has been described to be common to the N-terminal segments of the α1(V), α1,α2(IX), α1(XI) and α1(XII) chains as well as to a proline- and arginine-rich protein (PARP [31]; see Fig. 1). Multiple alignment and pattern searches indicate an additional similarity to the N-terminal part of thrombospondin (Table II, Fig. 2).

Thrombospondin is an adhesive mosaic protein containing three types of repeats [32] and a region which is similar to the Von Willebrand factor type C domain and to the N-terminal NC domains of several fibrillar procollagens [15]. No homology has been described so far for the N-terminal heparin binding domain of about 200 residues. This region exactly matches the module common to α1(V), α1,α2(IX), α1(XI), α1(XI), α1(XII) chains, and PARP (Fig. 3), even if most of the positions thought to be involved in heparin binding [32,33] are not conserved. The α1(V) and α1(XI) chains also contain clusters of positively charged residues in the putative heparin binding sites which have been proposed to be important in acidic proteoglycan binding [34]. The very remote relationship of collagen domains to throm-

Table II

Homology search (FASTA) statistics

| Module | Length of the domain | Rank of first false positive[1] | Number of probably right positives[2] | Maximal FASTA opt. score[3] | 5 best opt. scores[3] |
|---|---|---|---|---|---|
| Thrombospondin N-term. | 210 | 9 | 10 | 1057 | 136,119,118,118,111 |
| Collagen VII VWA | 200 | 20 | 24 | 1015 | 171,170,160,158,152 |
| Collagen VII Fn3 a | 90 | 54 | 240 | 449 | 124,119,119,118,117 |
| Collagen VII Fn3 b | 90 | 38 | 240 | 446 | 122,119,113,109,106 |

[1] The sequences detected by FASTA [48] were sorted for their optimized scores using FILTER_FASTA (Sander and Schneider, unpublished) which also filters the hits according to a threshold for structural homology [49]. Optimized FASTA scores include gap penalties and weights for similar amino acids.

[2] The number corresponds to the hits detected either by PAT or by PROFILESEARCH [50,51].

[3] The maximal possible opt. score is defined by self-comparison. The five best optimized scores in a database search with the respective domains all belong to proteins described in Table I and thus support the proposed relationships.

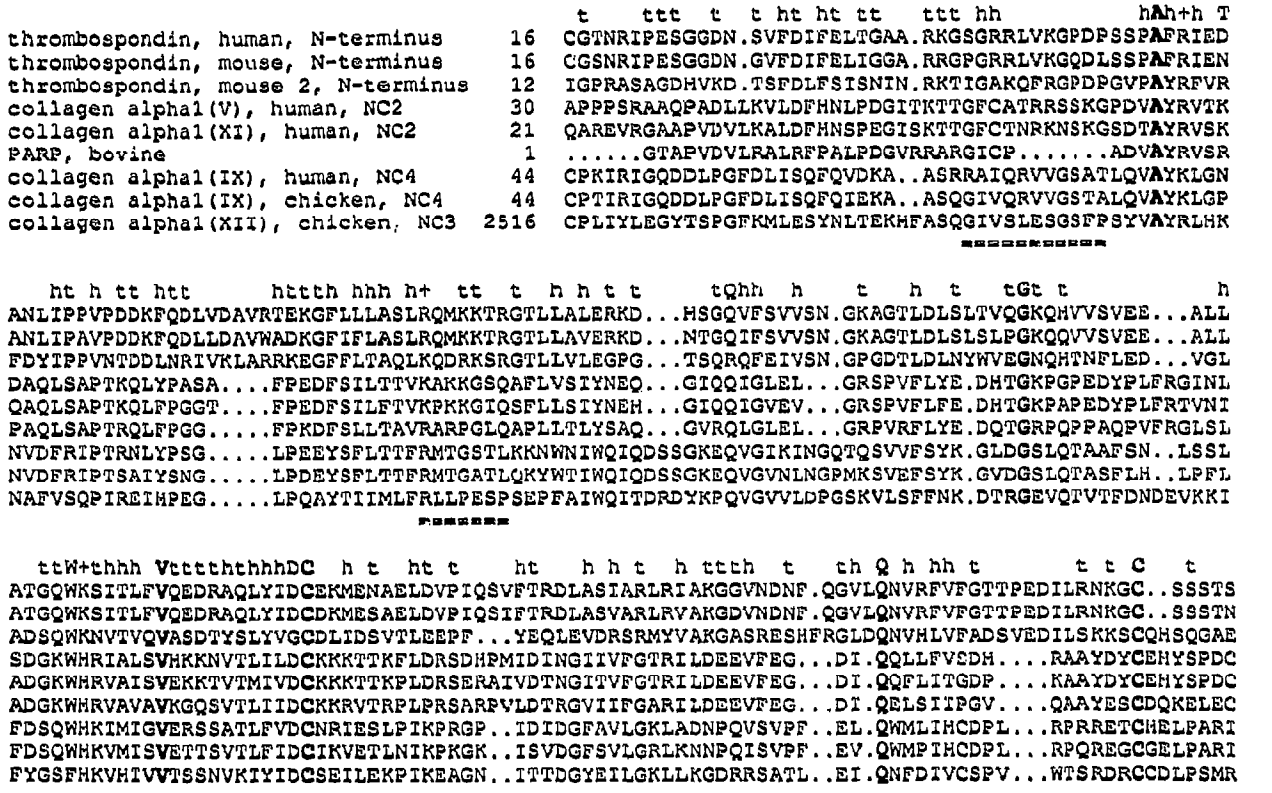


Fig. 2. Multiple alignment of the heparin binding domain in thrombospondins with NC domains of α1(V), α1,α2(IX), α2(IX), and α1(XII) chains as well as with PARP. The sequences are grouped into subfamilies according to their sequence similarity. The putative heparin binding sites in thrombospondin (underlined) are not conserved in the collagens, even if the α1(V) and α1(XI) chains also have clusters of positively charged residues in these regions. Conserved hydrophobic (h) and turn-like or polar positions (t) as well as invariant amino acids (tolerating one exception) are shown at the top. In addition to the high scores of the homology search program FASTA (Table II), two different pattern search tools (PROFILE and PAT) give significant hits for thrombospondins (PROFILESEARCH: z-score 14.03 with a random background of nonrelated sequences below a z-score of 6.55 and PAT: 3 mismatches by discrimination of noise effects appearing with more than 6 mismatches; for details see [50,51,65]).

bospondin is not surprising considering that this domain is the most divergent one between the two distinct thrombospondin genes [33]. These two genes also differ in their putative heparin binding sites [35].

Thrombospondin has been reported to bind specifically to numerous adhesive proteins (see e.g. [36]) including collagens [37,38]. Some of the responsible segments have already been characterized, but not yet mapped to the N terminus (see e.g. [39]). Thus, thrombospondin appears to have a complex regulatory role

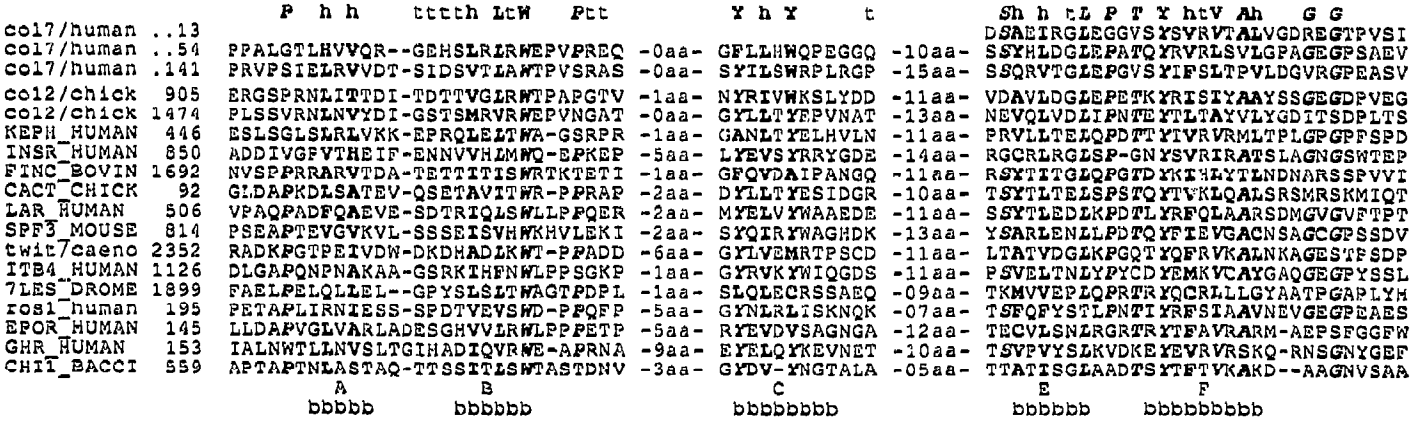


Fig. 3. Multiple alignment of type VII collagen sequences with selected fibronectin type III modules. SWISSPROT codes [62] are used when available. The consensus line notation is explained in Fig. 2. Fn3 consensus positions are shown in bold. The putative beta strands [67] are shown at the bottom. Note the similarity between the repeats of type VII collagen and those of type XII collagen.

within the extracellular matrix. The two regions in thrombospondin with similarity to distinct collagen types (Fig. 1, Table I) might be involved in anchoring different collagen types to specific tissues.

## 4. THE MODULAR ARCHITECTURE OF TYPE VII COLLAGEN

Sequence analysis indicates that two complete fibronectin type III (Fn3) modules, a part of a third one, and a Von Willebrand factor type A (VWA) domain constitute the N-terminal globular part of the human type VII collagen fragment (Table II, Fig. 3.)

Fn3 modules are found in a variety of adhesive proteins (Table I, for reviews see [5,10,11]). Comparative sequence analysis of more than 300 reported Fn3 modules (Bork and Doolittle, in press) reveals a relatively low overall similarity, but also several conserved features which are all present in type VII collagen (Fig. 2). Since the Fn3 module appears to have distinct functions in different groups of adhesive proteins (Bork and Doolittle, in press) no conclusion about a specific function can yet be drawn.

The VWA module is a domain of about 200 residues which has been reported so far in numerous extracellular adhesive mammalian proteins (Table I, [40,41]). All of these proteins appear to have a variety of binding functions.

In addition to type VII collagen, Fn3 repeats and VWA domains also coexist in type XII collagen [21], α3(VI) chain [22,23], and in undulin, another matrix protein without collagen-like triple helices [42]. Since type VII collagen is thought to have a large NC domain (e.g. [1,43]), additional repeats may be assumed. This, in turn, suggest a close similarity to the large NC4 domain of type XII collagen, supported by the relative high sequence identity between the Fn3 repeats in type VII collagen and those of type XII collagen as well as by the arrangement of Fn3 and VWA modules (Fig. 1).

In spite of possible functional switches of the domains in different proteins, a comparative analysis of globular modules in collagens gives some indication about common binding capacities and also highlights structurally important residues or regions (Figs. 2 and 3). The various similarities to other mosaic proteins and the presence of homologous modules in both fibrillar and nonfibrillar collagens suggests a consideration of NC domains in structural as well as functional classifications of collagens.

*Acknowledgements:* The author thanks C. Sander, R. Schneider, A. Valencia, and R. Wade for critical reading of the manuscript.

## REFERENCES

[1] Van der Rest, M. and Garrone, R. (1991) FASEB J. 5, 2814–2823.
[2] Gordon, M.K. and Olsen, B.R. (1990) Curr. Opin. Cell. Biol. 2, 833–838.
[3] Vuorio, E. and de Crombrugghe, B. (1990) Annu. Rev. Biochem. 59, 837–872.
[4] Engel, J. (1991) Int. J. Biol. Macrom. 13, 147–151.
[5] Bork, P. (1992) Curr. Opin. Struct. Biol. 2/3, 413–421.
[6] Van der Rest, M. and Garrone, R. (1990) Biochimie 72, 473–484.
[7] Banford, J.K.H. and Banford, D.H. (1990) Virology 177, 445–451.
[8] Gilbert, W. (1978) Nature 271, 501.
[9] Doolittle, R.F. (1992) Prot. Sci 1, 191–200.
[10] Patthy, L. (1991) Curr. Opin. Struct. Biol. 1, 351–361.
[11] Bork, P. (1991) FEBS Lett. 286, 47–54.
[12] Greenspan, D.S., Cheng, W. and Hoffman, G.G. (1991) J. Biol. Chem. 266, 24727–24733.
[13] Gordon, M.K., Gerecke, D.R. and Olsen, B.R. (1989) J. Biol. Chem. 264, 19772–19778.
[14] Yamaguchi, N., Benya, P.D., Van der Rest, M. and Ninomiya, Y. (1989) J. Biol. Chem. 264, 16022–16029.
[15] Hunt, L. and Barker, W.C. (1987) Biochem. Biophys. Res. Commun. 144, 876–882.
[16] Reid, K.B.M. and Day, A.J. (1990) Immunol. Today 11, 387–388.
[17] Sellar, G.C., Blake, D.J. and Reid, K.B.M. (1991) Biochem. J. 274, 481–490.
[18] Koller, E., Winterhalter, K.H. and Trueb, B. (1989) EMBO J. 8, 1073–1077.
[19] Chu, M.-L., Pan, T., Conway, D., Kuo, H.-J., Glanville, R.W., Timpl, R., Mann, K. and Deutzmann, R. (1989) EMBO J. 8, 1939–1946.
[20] Bonaldo, P., Russo, V., Bucciotti, F., Bressan, G.M. and Colombatti, A. (1989) J. Biol. Chem. 264, 5575–5580.
[21] Yamagata, M., Yamada, K.M., Yamada, S.S., Shinomura, T., Tanaka, H., Nishida, Y., Obara, M. and Kimata, K. (1991) J. Cell. Biol. 115, 209–221.
[22] Chu, M.-L., Zhang, R.-Z., Pan, T., Stokes, D., Conway, D., Kuo, H.-J., Glanville, R., Mayer, U., Mann, K., Deutzmann, R. and Timpl, R. (1990) EMBO J. 9, 385–393.
[23] Bonaldo, P., Russo, V., Bucciotti, F., Doliana, R. and Colombatti, A. (1990) Biochemistry 29, 1245–1254.
[24] Pihlajaniemi, T., Tryggvason, K., Meyers, J.C., Kurkinen, M., Lebo, R., Cheung, M.-C., Prockop, D.J. and Boyd, C.D. (1985) J. Biol. Chem. 260, 7681–7687.
[25] Babel, W. and Glanville, R.W. (1984) Eur. J. Biochem. 143, 545–556.
[26] Hostika, S.L. and Tryggvason, K. (1988) J. Biol. Chem. 263, 19488–19493.
[27] Morrison, K.E., Germino, G.G. and Reeders, S.T. (1991) J. Biol. Chem. 266, 34–39.
[28] Maryama, M., Kalluri, R., Hudson, B.G. and Reeders, S.T. (1992) J. Biol. Chem. 267, 1253–1258.
[29] Pihlajaniemi, T., Pohjolainen, E.-R. and Myers, J.C. (1990) J. Biol. Chem. 265, 13758–13766.
[30] Pettitt, J. and Kingston, I.B. (1991) J. Biol. Chem. 266, 16149–16156.
[31] Neame, P.J., Young, C.N. and Treep, J.T. (1990) J. Biol. Chem. 265, 20401–20408.
[32] Lawler, J. and Hynes, R.O. (1986) J. Cell. Biol. 103, 1635–1648.
[33] Lawler, J., Ferro, P. and Duquette, M. (1992) Biochemistry 31, 1173–1180.
[34] Vasios, G., Nishimura, I., Konomi, H. and van der Rest, M. (1988) J. Biol. Chem. 263, 2324–2329.
[35] Laherty, C.D., O'Rorke, K., Wolf, F.W., Katz, R., Seldin, M.F. and Dixit, V.M. (1992) Biochemistry 31, 3274–3281.
[36] Hogg, P.J., Stenflo, J. and Mosher, D.F. (1992) Biochemistry 31, 265–269.
[37] Lahav, J., Schwartz, M.A. and Hynes, R.O. (1982) Cell 31, 253–262.
[38] Galvin, N.J., Vance, P.M., Dixit, V.M., Fink, B. and Frazier, W.A. (1987) J. Cell. Biol. 104, 1413–1422.

[39] Dardik, R. and Lahav, J. (1991) Biochemistry 30, 9378–9386.
[40] Columbatti, A. and Bonaldo, P. (1991) Blood 77, 2305–2315.
[41] Bork, P. and Rohde, K. (1991) Biochem. J. 279, 208–210.
[42] Just, M., Herbst, H., Hummel, M., Dürkopf, H., Tripier, D., Stein, H. and Schuppan, D. (1991) J. Biol. Chem. 266, 17326–17332.
[43] Mayne, R. and Burgeson, R.E. (1987) Eds., Structure and function of collagens types, Academic press, Orlando, Florida.
[44] Kodama, T., Freeman, M., Rohrer, L., Zabrecky, J., Matsudaira, P. and Krieger, M. (1990) Nature 343, 531–535.
[45] Thiel, S. and Reid, K.B.M. (1989) FEBS Lett. 250, 78–84.
[46] Shimizu, H., Fisher, J.H., Papst, P., Benson, B., Lau, K., Mason, R.J. and Voelker, D.R. (1992) J. Biol. Chem. 267, 1853–1857.
[47] Charalambous, B.M., Keen, J.N. and McPherson, M.J. (1988) EMBO J. 7, 2903–2909.
[48] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 85, 3338–3342.
[49] Sander, C. and Schneider, R. (1991) Proteins 9, 56–68.
[50] Bork, P. and Grunwald, C. (1990) Eur. J. Biochem. 191, 347–358.
[51] Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Proc. Natl. Acad. Sci. USA 84, 4355–4358.
[52] Parente, M.G., Chung, L.C., Ryynänen, J., Woodley, D.T., Wynn, K.C., Bauer, E.A., Mattei, M.-G., Chu, M.-L. and Uitto, J. (1991) Proc. Natl. Acad. Sci. USA 88, 6931–6935.
[53] Lozano, G., Yoshifumi, N., Thompson, H. and Olsen, B.R. (1985) Proc. Natl. Acad. Sci. USA 82, 4050–4054.
[54] Muragaki, Y., Kimura, T., Ninomiya, Y. and Olsen, B.R. (1990) Eur. J. Biochem. 192, 703–708.
[55] Yoshioka, H. and Ramirez, F. (1990) J. Biol. Chem. 265, 6423–6426.
[56] Bernard, M.P., Chu, M.-L., Myers, J.C., Ramirez, F., Gilenberry, E.F. and Prockop, D.J. (1983) Biochemistry 22, 5213–5223.
[57] Su, M.W., Lee, B., Ramirez, F., Machado, M. and Horta, W. (1989) Nucleic Acids Res. 17, 9473–9473.
[58] Ala-Kokko, L., Kontusaari, S., Baldwin, C.T., Kuivaniemi, H., Prockop, D.J. (1989) Biochem. J. 260, 509–516.
[59] Woodbury, D., Benson-Chanda, V. and Ramirez, F. (1989) J. Biol. Chem. 264, 2735–2738.
[60] Pihlajaniemi, T. and Tamminen, M. (1990) J. Biol. Chem. 265, 16922–16928.
[61] Ninomiya, Y., Gordon, M., Van der Rest, M., Schmid, T., Linsenmayer, T. and Olsen, B.R. (1986) J. Biol. Chem. 261, 5050.
[62] Bairoch, A. and Boeckmann, B. (1991) Nucleic Acids Res. 19, 2247–2249.
[63] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[64] Bork, P., Sander, C. and Valencia, A. (1992) Proc. Natl. Acad. Sci. USA 89, in press.
[65] Bork, P. and Sander, C. (1992) FEBS Lett. 300, 237–240.
[66] Dublet, B. and van der Rest, M. (1991) J. Biol. Chem. 266, 6853–6858.
[67] Baron, M., Main, A.L., Driscoll, P.C., Mardon, H.J., Boyd, J. and Campbell, I.D. (1992) Biochemistry 31, 2068–2073.