Proc. Indian Acad. Sci. (Chem. Sci.), Vol. 111, No. 1, February 1999, pp. 105–113

Mechanisms of collagen trimer assembly

MICHEL VAN DER REST*, BERNARD DUBLET,
LIONEL LABOURDETTE and SYLVIE RICARD-BLUM
Institut de Biologie Structurale (CEA-CNRS), 41, Avenue des Martyrs, 38027 Grenoble Cedex 01, France
e-mail: mv@ibs.ibs.fr

Abstract. It is generally accepted that the folding of collagen triple helical domains occur from the C-terminus toward the N-terminus by a "zipper" mechanism. The regions at the C-terminus of the triple helices must therefore play a critical role in the processes of chain recognition and assembly to get the proper stoichiometries and of chain registration to align the chains for the folding of the triple helix. Examination of these regions reveals a broad diversity of structures and suggests that different mechanisms of assembly are used in the various collagen and collagen-like molecules. We review here three different mechanisms that have recently come to light. The collectins, a group of serum proteins containing collagen-like triple helical domains, are assembled through hydrophobic interactions in a triple α helix. Collagens VIII and X, C1q and several related proteins contain homologous C-terminal domains that are characterized by a β-pleated sheet structure. They assemble through very strong hydrophobic interactions that probably involve an "aromatic zipper". Collagens IX, XII and XIV fibril associated collagen with interrupted triple helices (FACITs), are assembled by a mechanism in which both the C-terminal triple helix and a very short cysteine-containing sequence are involved.

Keywords. Collagen; trimerization; disulfide bond; triple helix; protein folding.

1. Introduction

Collagens, the most abundant class of extracellular matrix proteins in animals, are characterized by the presence of one or more triple helical domains. The helical structure is made of polypeptide chains containing Gly-X-Y repeats, where X and Y typically represent proline or hydroxyproline residues. Due to the high iminoacid content of these repetitive sequences, each chain assumes a left-handed helical conformation that is similar to polyproline II helix. The three chains revolve around a common central axis and form a right-handed triple helix [1,2].

Assembly of the triple helix has been described for fibrillar collagens and involves a zipper-like C-terminal to N-terminal mechanism (for review see ref. 3) (figure 1). It has been shown that chains of purified C-propeptide of type I collagen self-associate to form trimers [4]. Although numerous studies have described the role of disulfide bonds in the

Abbreviations. FACIT – fibril associated collagen with interrupted triple helix; NC – non-triple helical domains; COL – triple helical domains; SMCD – schmid metaphyseal chondrodysplasia

* For correspondence

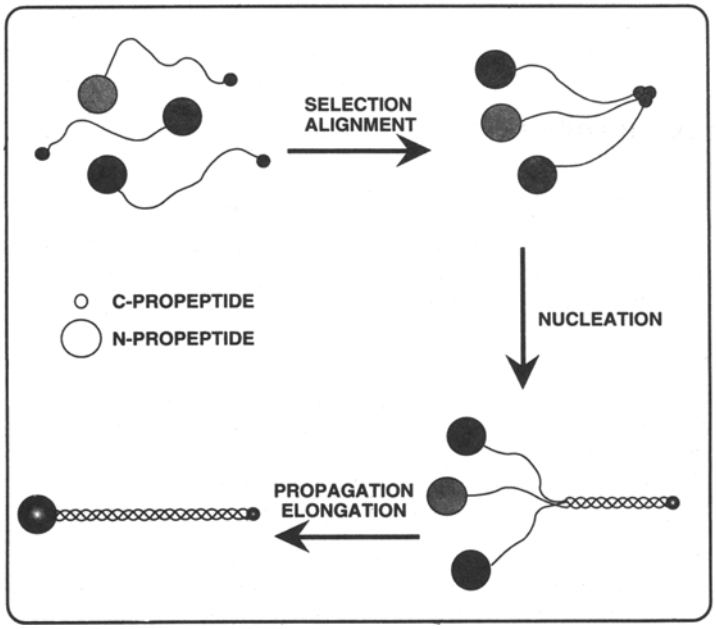

Figure 1. Major steps in the triple helical folding of collagens (adapted from Engel and Prockop [3]).

folding of the C-propeptides and in procollagen assembly, the initial steps of chain interaction and assembly have so far remained unresolved [5–8].

Nineteen different collagen types have now been reported (see reviews [9,10]), in addition to several other proteins that also contain collagen-like triple helices but are not considered collagens as they do not participate in the assembly of extracellular matrices. Such molecules include the C1q component of the complement, the Q subunit of acetylcholinesterase, the collectins, the scavenger receptor of macrophages. The same steps that have been outlined by Engel and Prockop [3] for the assembly and folding of fibrillar collagens (figure 1) must also take place for the other molecules containing triple helical collagen-like domains.

2. Directionality of triple helix folding

The folding of the triple helix in type I collagen occurs from the C-terminal end toward the N-terminus by the so-called "zipper" mechanism [3]. The most compelling evidence for this mechanism comes from the effects of mutations affecting the C-terminal propeptide or the triple helix that have been observed in Osteogenesis Imperfecta (OI), a heritable condition that leads to moderate to severe bone fragility. In one homozygous patient, a mutation in the C-propeptide of the pro α2(I) chain has been shown to prevent the integration of this chain into heterotrimeric molecules [11]. In a number of other OI patients with dominant mutations, substitutions of glycine in a Gly-X-Y triplet have been shown to increase the post-translational modifications only toward the N-terminus (upstream) from the mutations [12]. Since the enzymes involved are only active on the unfolded protein, this

shows that the triple helix downstream from the mutation must fold normally but not the region upstream.

This directionality of triple helix folding has not been as firmly proven for other collagen types. Although no theoretical argument against an N-terminal to C-terminal propagation of the helix has ever been presented, it is generally assumed that folding occurs from C-terminus toward N-terminus for all collagen-like triple helices. If this is correct, then the region C-terminal to triple helical domains must play a critical role in the processes of chain selection, assembly and registration for all these molecules.

3. Diversity of C-terminal domains

In view of the above considerations, one would also expect that the C-terminal domains, thought to be responsible for trimer formation, would be homologous between the molecules having collagen-like triple helical domains. This is not the case. Figure 2 presents a schematic comparison of the C-terminal domains of various collagen types. Based on the structure of these domains, one could classify the collagen family of proteins into several subgroups that would actually closely resemble the subgroups previously defined based on their overall structural and functional similarities[9]. The sizes of these domains as well as the number and the positions of the cysteines are very variable. If

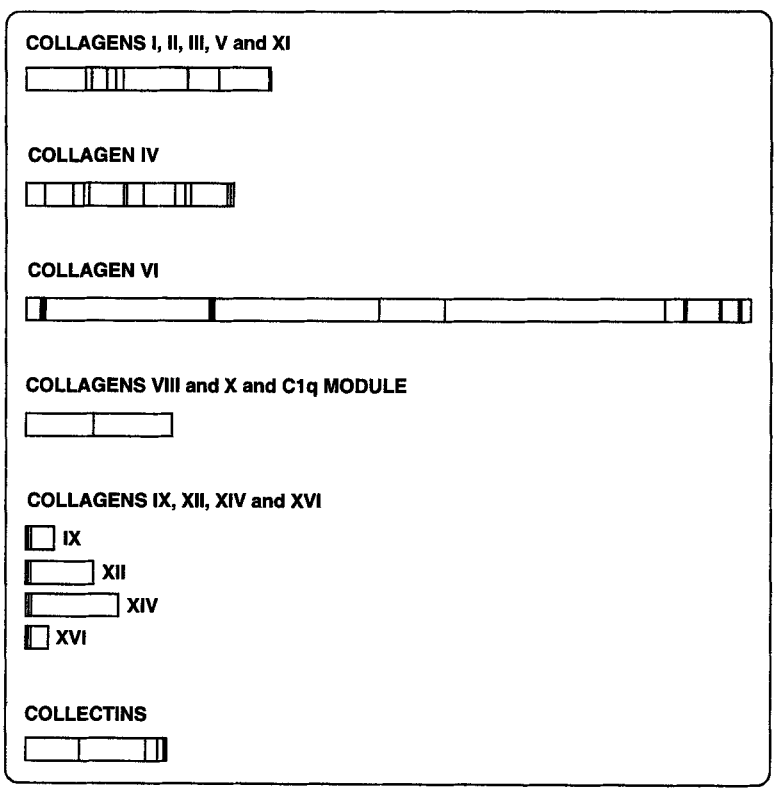

Figure 2. Comparison of the structures of the C-terminal domains of various collagen types. The lengths of the open boxes are proportional to the sizes of the domains. Vertical bars indicate the positions of cysteine residues.

these domains are indeed important for the trimerization process, then it is clear that Nature must have found quite different actors, and possibly mechanisms, to assemble collagen and collagen-like molecules. We discuss here some of the recent data that have been obtained in our and other laboratories and attempt to throw some light on the diversity of trimerization processes in the various collagens.

4. Assembly through amphipathic α helices

The best described assembly mechanism for a molecule containing collagen-like triple helices is the one found in the collectins, a group of molecules characterized by the presence of a collagen-like triple helical domain and a C-type lectin domain [13]. These molecules include the serum proteins mannan-binding protein, collectin-43, conglutinin and lung surfactant proteins SP-D and SP-A (figure 3). They are characterized by the presence of amphipathic α helices located in the C-terminus from the triple helical domain. By hydrophobic interactions between leucine and isoleucine residues that are located on one edge of these helices, a triple α helix is formed that serves as a recognition site for the constituent chains and initiates trimerization. It should be noted that a globular C-type lectin domain is located further toward the C-terminus in the collectin chains. It does not appear to be involved in the assembly process. Assembly of collagen and collagen-like molecules does not therefore necessarily have to be initiated from the very C-terminal end of the molecule but domains irrelevant to this process may be located between the assembly site and the C-terminus of the chains.

5. Assembly through globular domains with β-pleated sheets

A conformation analysis performed in 1992 by Brass *et al* [14], and based on Fourier transform infrared spectroscopy and on structure prediction from molecular modeling, suggested that the C-terminal propeptides of fibrillar collagens and the NC1 C-terminal domains of collagens VIII and X, of C1q and of several homologous trimeric proteins are folded into β-pleated sheets. A short sequence with conserved aromatic amino acids found in all these domains was proposed to play an essential role in the assembly process. It has often been referred to as the "aromatic zipper". These conclusions have been confirmed by the determination of the three-dimensional structure of the C-terminal trimeric domain of the adipocyte complement-related protein of 30 kDa (AdipoQ), a domain closely homologous to the NC1 domains of collagens VIII and X and to the C-terminal globular heads of C1q [15]. The presence of 10 β-strands has been confirmed. Interestingly, the aromatic zipper is located on a single strand in the hydrophobic core of the structure where the three subunits interface with each other. This observation goes along the role proposed for this sequence.

6. Effects of mutations in the NC1 domain of collagen X on assembly

In our laboratory, we have recently investigated the effect of mutations found in Schmid metaphyseal chondrodysplasia (SMCD) patients on the folding of the NC1 domain of collagen X (B Dublet, T Vernet and M van der Rest, submitted for publication). This autosomal dominant condition, characterized by a short stature and a waddling gait, has indeed been linked to the COL10A1 gene, coding for the α1(X) collagen chain. Missense and nonsense mutations have been described in the region of the gene coding for the NC1

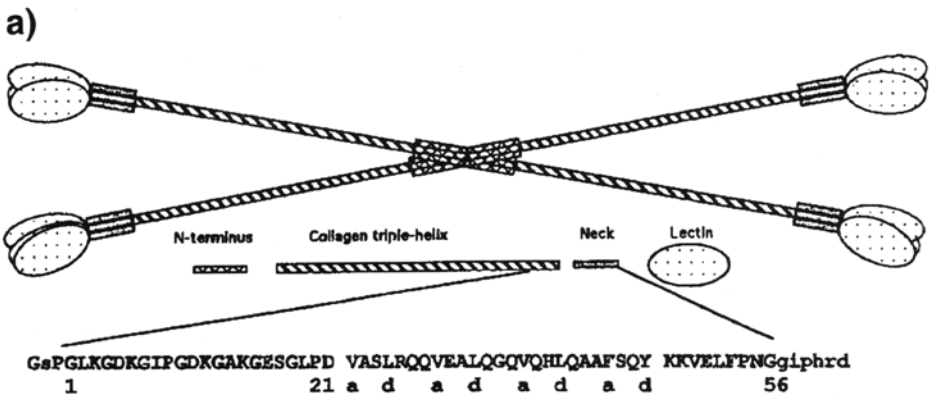

Figure 3. (**a**) Structure of lung surfactant SP-D, a model of collectins. It consists of 12 identical polypeptide chains assembled into 4 rod-like structures made of collagen-like triple helices. The C-terminal C-type lectin domains are represented by globules and are connected to the triple helical regions by the so-called neck regions which consist of triple α helices and are responsible for trimer formation. The N-terminal region is responsible for tetramerization. (**b**) Projection of residues 22–48 of the neck region onto a helical wheel. (Reproduced from Hoppe *et al* [13] with permission.)

domain, suggesting that the mutations interfere with the assembly of the molecule [16–18]. Our data have demonstrated that a recombinant protein expressed in *E. coli* and comprising the last 136 amino acids of the α1(X) chain (residues 545 to 680), corresponding to a C1q module as defined by Bork [19], does spontaneously trimerize. The trimer formed is extremely stable to denaturation and proteolytic cleavage. This confirmed previous reports on the role of this domain in the assembly process of collagen X and its remarkable stability [20,21].

Four constructs were made in which four mutations found in SMCD were separately introduced. Two missense mutations were made, one involving a Y residue of the aromatic zipper (Y598D) and the other 21 positions downstream (G618V). Two premature termination codons were also introduced at positions coding for residues 651 and 659 respectively. All our constructs were made as gluthathione S-transferase fusion proteins and could thus be purified by glutathione sepharose chromatography. When the mutant proteins were purified, we observed that a major protein from *E. coli* copurified with them. This protein was identified by N-terminal sequencing as the GroEL bacterial chaperone protein. No GroEL was found associated with the normal protein when purified in the same way. In addition, all four mutant proteins had lost the resistance to proteolysis observed for the normal protein.

From these observations, one can conclude that SMCD mutations interfere with the normal folding process of the NC1 domain of collagen X rather than with the trimerization process itself. If proteins mutated in the NC1 domain are also sequestered by a chaperone protein in hypertrophic chondrocytes that synthesize collagen X, this could explain the relatively uniform phenotype observed in SMCD patients despite the diversity of the mutations and suggest that the phenotype is caused by haploinsufficiency.

7. Mechanism of assembly of fibril-associated collagens with interrupted triple helix (FACITs)

The initial stages of chain assembly must differ in the case of the FACITs since the sizes of their non-triple helical C-terminal (NC1) domains are relatively limited (15 to 119 residues) and present no homology with any of the C-terminal domains discussed above [10]. The primary sequences of the NC1 domains of FACITs do not contain any consensus sequence except for the first 5 residues (see below). These collagens however display a remarkable similarity in their C-terminal triple helical domain (COL1), and at the junction with the C-terminal globular domain (NC1) where two cysteines spaced by 4 amino acids and involved in interchain bonds are strictly conserved. We have thus tested the hypothesis that this region may play a role in the initial steps of chain assembly of FACIT molecules. Several approaches were taken.

In the first approach, we expressed a recombinant "mini" collagen XII in HeLa cells. The construct, placed after the cytomegalovirus (CMV) promoter, contained the sequences coding for the C-terminal triple helical (COL1) and non-triple helical (NC1) domains. The results showed that trimer formation, as evidenced by disulfide bond formation, was dependent on prolyl hydroxylation [22]. These data suggested that, in the case of FACIT molecules, the triple helical region was involved in the assembly process. Further data showed that deletion of most of the NC1 domain, except for the cysteine containing the first few residues, did not interfere in this system with trimer formation [23]. Similar experiments on heterotrimeric collagen IX where three "mini" chains [α1(IX), α2(IX), α3(IX)] were expressed separately and simultaneously in HeLa cells, showed that

heterotrimeric 1:1:1 assembly was favored, but that, in addition, depending on the ratios of the expressed chains, other non-natural stoichiometries could also be obtained[24].

A detailed analysis using pepsin fragments of the natural collagen IX protein covering the COL1 and part of the NC1 regions as well as synthetic peptides covering the COL1-NC1 junction was then performed[25,26]. Experiments with synthetic peptides were also independently reported by Mechling *et al*[27]. These data initially proved to be difficult to interpret and to reconcile with the results of the overexpression experiments. We have developed a model (figure 4) for the assembly of FACIT molecules. In this model, based on the experimental evidence that full folding of the COL1 triple helix only occurs when disulfide bonds are formed, we propose that the initial step of chain selection and assembly involves both the short cysteine-containing region of NC1 and the folding of an adjacent portion of the triple helix (step 1 on figure 4). This initial assembly would then be stabilized by disulfide bond formation (step 2) followed by the extension of the folding of the triple helix toward the N-terminus (step 3). This model implies that the "zipper" mechanism does not proceed at the same speed all along the COL1 domain. The presence of conserved imperfections in the repetitive Gly-X-Y sequence might be the basis of such an uneven rate of helix folding. *In vitro*, the predominant presence of monomeric and dimeric molecules with oxidized cystine residues indicates that most of the disulfide bonds formed are non-natural and correspond to intrachain bonds or result in the stabilization of dimers (step 4). In addition, circular dichroism analysis indicates the reversible formation of triple helical structures that do not form disulfide bonds, probably because of mismatches in the registration of the chains (step 5). If non-natural disulfide bridges do not trap the chains, this mechanism of "trials and errors" could proceed until the proper registration of the chains is obtained.

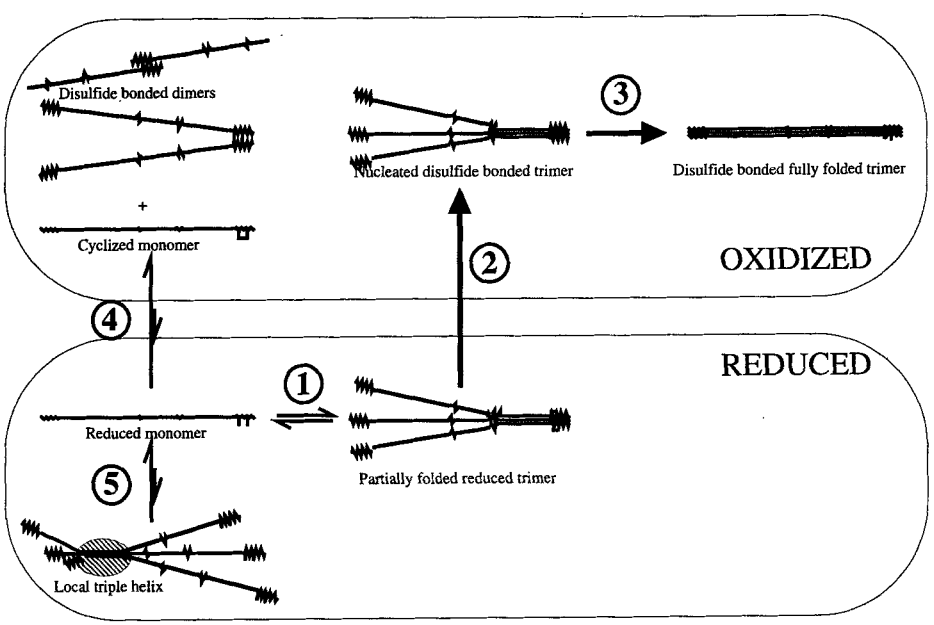

Figure 4. Proposed mechanism for the assembly of FACIT molecules.

In vivo, it is likely that an enzyme such as the protein disulfide isomerase (PDI) could play a critical role in preventing the unproductive pathway (step 4) observed *in vitro*. This model reconciles the role for the triple helical region demonstrated by the experiments of overexpression of the "mini" collagens with the role of the cysteine containing sequence in NC1 underlined by Mechling *et al*[27].

In this model, the folding of part of the triple helix is necessary to initiate trimerization while the cysteine-containing region would provide the specificity for proper chain registration and permit the permanent stabilization of the assembly. A three-dimensional structure of the COL1-NC1 junction has been obtained by NMR spectroscopy for a homotrimeric molecule made of three 19 amino acid-long synthetic peptides comprising three Gly-Pro-OHPro triplets followed by the cysteine-containing sequence of collagen XIV[28]. The asymmetry of the structure obtained gives clues as to how this region could contribute to chain selection and registration.

8. Conclusion

We have discussed here three different mechanisms used for assembly of collagen or collagen-like molecules. It is quite possible that other mechanisms are used for other members of the collagen family of proteins with different C-terminal structures. It also remains possible that some molecules could assemble and fold from the N-terminus, but no experimental evidence for such a reverse triple helix folding has been reported.

The diversity of mechanisms discussed here does not modify the basic principles presented by Engel and Prockop[3] for collagen folding. Chain selection and alignment occurs in all cases at the C-terminus but the structural motifs used for these steps are different. In the cases of collectins and C1q modules, hydrophobic forces appear to play the major role in the process, but in one case the structures are α helical while for the other case they are made mainly of β-pleated sheets. In the case of FACITs, the formation of the collagen triple helix itself, guided by a short cysteine containing sequence, plays the central role. It thus happens that the three main secondary structures defined by Ramachandran[29] in his famous diagram can all be used to assemble collagen and collagen-like molecules into trimers.

Acknowledgements

The work reported here has been supported by the Centre National de la Recherche Scientifique (CNRS), the Commissariat à l'Énergie Atomique (CEA) and the Biotech program of the European Commission.

References

1. Ramachandran G N and Ramakrishnan C 1976 In *Biochemistry of Collagen* (eds) G N Ramachandran and A H Reddi (New York: Plenum) pp. 45–84
2. Kühn K 1987 In *Structure and function of collagen types* (eds) R Mayne and R E Burgeson (Orlando, FL: Academic Press, Inc.) pp. 1–42
3. Engel J and Prockop D J 1991 *Annu. Rev. Biophys. Biophys. Chem.* **20** 137
4. Doege K J and Fessler J H 1986 *J. Biol. Chem.* **261** 8924
5. Fessler L I, Burgeson R E, Morris N P and Fessler J H 1973 *Proc. Natl. Acad. Sci. USA* **70** 2993
6. Byers P H, Click E M, Harper E and Bornstein P 1975 *Proc. Natl. Acad. Sci. USA* **72** 3009

7. Harwood R, Merry A H, Woolley D E, Grant M E and Jackson D S 1977 *Biochem. J.* **161** 405
8. Gerard S, Puett D and Mitchell W M 1981 *Biochemistry* **20** 1857
9. van der Rest M and Garrone R 1991 *FASEB J.* **5** 2614
10. Mayne R and Brewton R G 1993 *Curr. Opin. Cell Biol.* **5** 883
11. Deak S B, Nicholls A, Pope F M and Prockop D J 1983 *J. Biol. Chem.* **258** 15192
12. Wenstrup R J, Shrago-Howe A W, Lever L W, Phillips C L, Byers P H and Cohn D H 1991 *J. Biol. Chem.* **266** 2590
13. Hoppe H J, Barlow P N and Reid K B M 1994 *FEBS Lett.* **344** 191
14. Brass A, Kadler K E, Thomas J T, Grant M E and Boot-Handford R P 1992 *FEBS Lett.* **198** 208
15. Shapiro L and Scherer P E 1998 *Curr. Biol.* **8** 335
16. Warman M L, Abbott M, Apte S S, Hefferon T, McIntosh I, Cohn D H, Hecht J T, Olsen B R and Francomano C A 1993 *Nature Genet.* **5** 79
17. Wallis G A, Rash B, Sweetman W A, Thomas J T, Super M, Evans G, Grant M E and Boot-Handford R P 1994 *Am. J. Hum. Genet.* **54** 169
18. McIntosh I, Abbott M H and Francomano C A 1995 *Hum. Mutat.* **5** 121
19. Bork P and Koonin E V 1996 *Curr. Opin. Struct. Biol.* **6** 366
20. Schmid T M and Linsenmayer T F 1984 *Biochemistry* **23** 553
21. Barber R E and Kwan A P 1996 *Biochem. J.* **320** 479
22. Mazzorana M, Gruffat H, Sergeant A and van der Rest M 1993 *J. Biol. Chem.* **268** 3029
23. Mazzorana M, Giry-Lozinguez C and van der Rest M 1995 *Matrix Biol.* **14** 583
24. Giry-Lozinguez C 1996 *Functional study of FACIT collagens through expression of recombinant proteins*, Ph D thesis, University of Lyon, Lyon
25. Labourdette L and van der Rest M 1993 *FEBS Lett.* **320** 211
26. Labourdette L 1994 *Molecular interaction mechanisms of type IX collagen*, Ph D thesis, University of Lyon, Lyon
27. Mechling D E, Gambee J E, Morris N P, Sakai L Y, Keene D R, Mayne R and Bächinger H P 1996 *J. Biol. Chem.* **271** 13781
28. Lesage A, Penin F, Geourjon C, Marion D and van der Rest M 1996 *Biochemistry* **35** 9647
29. Ramachandran G N and Sasisekharan V 1968 *Adv. Protein Chem.* **23** 283