Proc. Indian Acad. Sci. (Chem. Sci.), Vol. 111, No. 1, February 1999, pp. 81–86

Collagen-like sequences in phages and bacteria

JUERGEN ENGEL*[1] and HANS PETER BÄCHINGER[2]
[1]Department of Biophysical Chemistry, Biozentrum, University of Basel, Klingelbergstr. 70, CH-4056 Basel, Switzerland
[2]Research Unit, Shriners Hospital for Children and the Department of Biochemistry and Molecular Biology, Oregon Health Science University, Portland, Oregon, 97201, USA
e-mail: engel@ubaclu.unibas.ch; HPB@shcc.org

Abstract. Sequences with glycine in every third position have been detected in DNA derived sequences of proteins in phages and bacteria and it was suggested that these regions trimerise to collagenous structures. Related sequences are found in proteins of mollusks and slime mold. The sequences contain a much lower fraction of proline than mammalian collagens and it is unknown how many of the prolines, if any, are converted to hydroxyproline. Therefore, for triple helix formation other stabilizing interactions than those known for mammalian collagens are required. Strikingly, aspartate and asparagine are abundant in collagen-like sequences of phage tail fibre proteins and of related sequences in nacrein of oyster pearls suggesting a possible stabilization by calcium binding.

Keywords. Triple helix; stabilization; evolution.

1. Introduction

In the past, research on collagens was focused on mammalian collagens of which 19 members, collagen I to XIX, are known today [1,2]. This classification does not include collagens which are not directly connected to the extracellular matrix, like C1q and the collectins[3]. All collagens contain by definition collagen domains but also a large variety of other domains. The number of genes is higher than the number of collagens because each collagen consists of three chains, which are often different. For the formation of the collagen triple helix the occurrence of glycine in every third position is a prerequisite, because larger side chains than hydrogen atoms would interfere with the packing of three polyproline helices to the triple helix [1,4]. Therefore collagenous sequences can be easily recognized by their Gly–Xaa–Yaa repeat. Another characteristic feature of most collagenous sequences is the occurrence of hydroxyproline in Y-positions. It was recognized that this residue is essential for the stabilization of the triple helix. In addition proline is very abundant and also exhibits a stabilizing effect on the polyproline type II helices. The denaturation temperature Tm of collagens is near to the living temperature of the organisms [5] and for many collagens a monotonous relation between Tm and the sum of proline and hydroxyproline is observed[6].

Many collagens of lower organisms are homologous to mammalian collagens[7]. Examples are the fibril forming collagen I from annelids [8] and the basement membrane

* For correspondence

collagen IV from *Drosophila*[9] and *C. elegans*[10]. On the other hand several collagens were found in invertebrates for which no resemblence exists with mammalians[7]. The byssus collagen of mussel[11], the cuticle collagens of worms[8] and the mini-collagen of nematocytes in hydra[12] may serve as examples. Very interestingly, new ways of triple helix stabilization have been found for invertebrate collagens. In an annelid cuticle collagen, glycosylated threonines in Y-position replace the normally occuring hydroxyprolines[13]. By studies with model peptides it was found that glycosylated threonines stabilize the triple helical structure[14].

Very recently, the presence of collagens in several lower organisms including phages and bacteria was proposed on the basis of cDNA sequences or genomic analysis[15–17]. In these cases usually no information exists about post-translational modifications like hydroxylation or glycosylation. Evidence is missing whether triple helices are formed from the Gly–Xaa–Yaa repeat sequences or not. In the present work a number of these sequences are compared and their unusual amino acid compositions are discussed from the point of view of triple helix stabilization.

2. Results and discussion

In table 1 several recently discovered collagen-like sequences of bacteriophages are summarized together with similar sequences of other organisms. All sequences were collected from the protein data base TrEMBL (transl.EMBL), release 5.0 (2/98) and accession numbers are included to facilitate a search for further sequence information. The sequences of the phage proteins originate from genome sequencing and are frequently obtained by a combination of the sequences of several open reading frames. Sequence information on the protein level does not exist and it is therefore not known whether or not prolines are converted to hydroxyprolines or whether these or other residues are glycosylated or otherwise post-tranlationally modified. It may however be argued that phages and bacteria do not process hydroxylation and suitable glycosylation systems in their genome and that such modifications are therefore unlikely. Genetic and biochemical information is still rather incomplete but it was suggested that the collagen-like proteins are localized in the phage tail fibres[16,17]. An exception is a protein from bacteriophage PRD1 which was localized in the phage head[15]. It has a repeat of six GXY residues, which was considered a nucleation site for trimerization of this oligomeric protein.

It is not known from biochemical or biophysical data whether the putative tail fibre proteins can form homo- or hetero-trimers by triple helix formation of their collagen-like regions and what the denaturation temperatures Tm are. a prediction of triple helix stability by comparison with mammalian collagens cannot be made becuse of the unusual composition of residues in X and Y positions. Mammalian collagens typically contain more than 10 mol% proline and an even higher fraction of hydroxyproline, together more than 20 mol%. In the collagenous translated sequences of the tail fibre phage BK-5T protein as an example only 6·7 mol% proline are found. A striking feature of the Gly–Xaa–Yaa repeats in the phage proteins is the unusually high fraction of aspartate (D) and asparagine (N) which together constitute 48% of all residues in X and Y positions in phage BK-5T (table 1). Interstingly a data base scan with D- and N-rich sequence motifs revealed a number of similar sequences in bacteria, mollusks, fungi and slime mold of which examples are shown in table 1. GNN and GDN repeats are very frequent and the total fraction of D and N residues in X- and Y-position amounts to 83% in nacrein from

Table 1 Collagen-like sequence regions in phages, viruses, bacteria, mollusks, fungi and slime mold.

q38319 lactococcal phage BK-5T putative tail fiber protein		
	0848	GNDGKDGATGKDGVAGKDGVG
	0940	GNNGNDGIAGKDGVG
	1003	GVKGDKGDPGNNGTNGIAGKD
	1084	GTNGNNGHDGFPGKDGTG
	1159	GVKGDKGDPGNNGTNGIAGKDGKDGKG
	1240	GTNGNNGHDGFPGKDGTG
	1315	GVKGDKGDPGNNGTNGIAGKDGKG
	1396	GTNGNNGHDGFPGKDGTG
	1471	GVKGDKGDPGNNGTNGIAGKDGKG
	1552	GTNGNNGHDGFPGKDGTG
	1624	GKMGNTGPAGSNGNPGKV
o34076 bacteriophage ϕ01205 putative tail fiber protein		
	0190	GAAGPKGDQGNDGLPGKDGVG
	0270	GNNGNDGLPGKDGVG
	0342	GEQGPKGDRGRQGLQGPRGEQGIPGPKGADGRT
	0424	GSDGKDGVPGKAGADGRT
z50114 coliphage BF23 putative tail fiber protein		
	0123	GVDGRPGADGKPGADGKPGADGRPGDNGQRGPG
q69475 herpesvirus putative nuclear antigen		
	0183	GDDGDDGDEGGDGDEGEEGQE
q98182 molluscum contagiosum virus hypothetical protein		
	0184	GDDGGDGGNGGNGGDGGDGGD
o06810 mycobacterium 61 kD protein		
	0501	GADGTDGGKGGNGGAGGG
	O819	GGDGGDGGNGGN
	0845	GNGGDGGNGGNGGSAGTGGNGGRGGDG
	0875	GRNGPNGNPGGNGGAGGAGGAGLNGGNGGGAGGNGGLGGFGGN
	1030	GGNGGHGGHGAA
	1043	GGNGGPGGHGGNGGNGGTGANGGNGGIGGTGGAGSTGAKGVLGTN
q52544 pseudomonas POPA 1 protein		
	0186	GVGGAGGADGGSGAGGAGGANGADGGNGVNGNQ
q27908 oyster pearl nacrein		
	0242	GDNGNNGYNGDNGNNGDNGNN
	0266	GDNGNNGVNGNNGYNGDNGNNGDNGNNGYNGDNGNNGDNGNNGEN GNNGENGNNGENGHK
q09164 tolypocladium inflatum cyclosporin synthase		
	15179	GTNGTNGTNGTNGANGTNGTNGTNGTH
p90535 dyctyostelium discoidium RSC12 fragment		
	0390	GNNGNNGNNGNNGNNGNNGNNGNNGNN
	0425	GNNGNNGNNGNNGNNGNNGNNGNNGNNGNNGNNGQS

oyster pearls and almost 100% in the dyctyostelium discoidium RSC12 fragment (table 1). Aspartate and asparagine as well as threonine which also occurs in the N and D-rich sequences are frequently involved in complex formation with Ca^{2+} or other bivalent ions[18]. It may therefore be hypothesized that the lack of proline and hydroxyproline is balanced by a stabilization resulting from metal ion complexation to N and D rich triple helices. Model building shows ligandation of the ions by side chains of Asp and Asn might be possible in a triple helical conformation.

Data on calcium binding so far only exist for nacrein, which is a carbonic anhydrase from the nacreous layer in oyster pearls. For this protein it was demonstrated by a

Table 2. Domains with collagen-like sequences in putative tail fiber proteins of bacteriophages.

q38319 lactococcal phage BK5-T

0848	GNDGKDGATGKDGVAGKDGV	GIKTTVITYALSSSGTDKPNTGWTSQVPTLVKGQYLWTKT VWTYTDSSSETGYSVTYIAKD
0940	GNNGNDGIAGKDGV	GIK KTTITYAVGTSGTTAPA SGWNSQVPNVPAGQFLWTKTVWTYTDNTSETGYSVAMM
1003	GVKGDKGDPGNNGTNGIAGKD	GIKATAITYQASPNGTTAPTGTWSASVPPVAKGSFLWTRTIWTYTDNTTETGYAVAYM
1084	GTNGNNGHDGFPGKDGTGIK	TTTITYAGSTSGTTPPNNGWTSTVPTVAEGNYLWTKTVWTYTDNTSETGYSVAMM
1159	GVKGDKGDPGNNGTNGIAGKDGKG	IKATAITYQASPNGTTAPTGTWSASVPPVAKGSFLWTRTIWTYTDNTTET
1240	GTNGNNGHDGFPGKDGTGIK	TTTITYAGSTSGTTPPNNGWTSTVPTVAEGNYLWTKTVWTYTDNTSETGYSVAMM
1315	GVKGDKGDPGNNGTNGIAGKDGKG	IKATAITYQASPNGTTAPTGTWSASVPPVAKGSFLWTRTIWTYTDNTTETGYAVAYM
1396	GTNGNNGHDGFPGKDGTGIK	TTTITYAGSTSGTTPPNNGWTSTVPTVAEGNYLWTKTVWTYTDNTSETGYSVAMM
1471	GVKGDKGDPGNNGTNGIAGKDGKG	IKATAITYQASPNGTTAPTGTWSASVPPVAKGSFLWTRTIWTYTDNTTETGYAVAYM
1552	GTNGNNGHDGFPGKDGTGIK	TTTITYAGSTSGTTPPNNGWTSTVPTVAEGNYLWTKTVWAYTDNSFETGYSV
1624	GKMGNTGPAGSNGNPGKV	VSDTEPTTKFKGLTWKYSGVVDMPLGNGTKILAGTEYYWNGNNWALYEI

o34076 bacteriophage ϕ01205

0190	GAAGPKGDQGNDGLPGKDGVG	IKTTIVTYGISDNENTQPTNWSSQLPTLVKGKYLWTKTAWTYTDLSSETGYQKTYIAKD
0270	GNNGNDGLPGKDGV GIR	NTTITYAVGTSGTVAPTNGWSSQVPNVPAGQYLWTKSIWDYTDNTSETGYSVAKM
0342	GEQGPKGDRGRQGLQGPRGEQGIPGPKGADGRTQYTHIAYADAISGSGFSQTDVSKPYIGMYQDFNE VDSNNPQDYRWSKWK	
0424	GSDGKDGVPGKAGADGRTPYVHFAYADSADGRTGFSLTQNGRKRYLGVLTNFIKKDSTNPSDYSWNDTAGSVSVGGENLIRNSAFPKNLDGWGHW	

z50114 coliphage BF23

0123	GVDGRPGADGKPGADGKPGADGRPGDNGQRGPGMYSLAIANLTAWNDSQANAFFTSNFGTGPVKYDVLTEYKSGAPGTAFTRQWNGSAWT

qualitative assay that Ca^{2+} binds to the region containing Gly–Xaa–Asn repeats with Xaa frequently being apartate[19]. However, our preliminary circular dichroism spectra of a synthetic peptide comprising of residues 266 to 326 in nacrein did not indicate triple helix formation in the presence or absence of calcium. A low potential for forming a triple helix was also found for synthetic pepetides in which the sequence Gly–Asn–Asn is repeated ten times. Experiments with these and related peptides will be continued introducing nucleation and cross-linking sites for promotion of triple helix formation. In the systems listed in tables 1 and 2 additional triple helix stability may originate from the adjacent globular domains. Furthermore, individual triple helices might assemble to superstructure by hydrogen bonding between glycine residues in different helices. Such supermolecular assemblies have been demonstrated for model peptides containig adjacent glycine residues. Examples are the sheet-like structures of synthetic peptides with Gly–Gly–Pro repeats and the three-dimensional assembly of polyglycine[4,20].

In table 1 uninterrupted collagenous sequences are shown at their full length and all the regions following after an interruption are indicated. It can be seen that interruptions are of very variable length with the exception of the phage tail fibre proteins. They exhibit a repeat structure in which each collagen-like sequence is followed by a non-collagenous region of about 50 residues. These regions according to their amino acid composition probably form small globular domains. Within a single protein for example of phage BK-5T a clear internal homology is detectable for the 50 residue repeats (table 2). Homologous domains of this type were also found in tail proteins of phages Dp-1 and ϕSfi21[17]. A long stretch of 11 collagen domains, each followed by a globular domain of 50 residues is predicted with the assumption that each collagen-like sequence region forms a triple helix. The estimated length of this region in BK-5T is about 100 nm assuming a 0·3 nm translation for residues in the triple helix and a diameter of 2·5 nm per globular domain. Extended shapes are expected since tail fibres are very long (160 nm in T4 phage[21]). It should be mentioned that not all phages contain tail fibre proteins with a collagen-like sequence. The well studied T4 phages contain tail proteins with repeating cross β-structures. These proteins are also three-stranded[22].

The repeat structure in the putative collagenous phage tail proteins were probably created by gene duplications. Repeats emerging from gene duplications were also discussed for mammalian collagens[23,24] but are more difficult to prove because of the low preservation of collagen sequences not considering the glycines which repeat in every third position in all collagens. In the tail protein of phage BK5T conservation between collagenous sequences is also low but identity between the globular regions of the eleven repeats is higher than 50%. This high identity suggests that the repeats are relatively recent and argues against the hypothesis that these collagens of phages are the ancestors of mammalian collagens[16]. For phages and bacteria, horizontal gene transfer is a likely mode of adopting a gene[24]. This may also be the reason why related collagen-like sequence regions are found in bacteria which are the hosts of the bacteriophages. Studies of collagens in bacteriophages are still at a very early stage. Further work will be important for a better understanding of the evolution of collagens but may also provide new insights in the stabilization of collagen triple helices.

Acknowledgement

This work was supported by a grant to JE by the Swiss National Science Foundation.

References

1. Bateman J F, Lamandé S R and Ramshaw J A M 1996 In *Collagen family in extracellular matrix* (ed.) W D Comper (Amsterdam: Harwood) vol. 2 pp. 22–67
2. van der Rest M and Garrone R 1991 *FASEB J.* **5** 2814
3. Hoppe H J and Reid K B 1994 *Protein Sci.* **8** 143
4. Traub W and Piez K A 1971 *Adv. Protein Chemistry* **25** 243
5. Rigby B J 1968 *Nature (London)* **219** 166
6. Josse J and Harrington W F 1964 *J. Mol. Biol.* **9** 269
7. Engel J 1997 *Science* **277** 1785
8. Gaill F, Wiedemann H, Mann K, Kühn K, Timpl R and Engel J 1991 *J. Mol. Biol.* **221** 209
9. Blumberg B, MacKrell A J and Fessler J H 1988 *J. Biol. Chem.* **263** 18328
10. Kramer J M 1994 *FASEB J.* **8** 329
11. Coyne K J, Qin X X and Waite J H 1997 *Science* **277** 1830
12. Kurz E M, Holstein T W, Petri B M, Engel J and David C N 1991 *J. Cell Biol.* **115** 1159
13. Mann K, Mechling D E, Bächinger H P, Eckerskorn C, Gaill F and Timpl R 1996 *J. Mol. Biol.* **16** 255
14. Bann J E and Bächinger H P (unpublished results)
15. Bamford J K H and Bamford D H 1990 *Virology* **177** 445
16. Smith M C M, Burns N, Sayers J R, Sorrell J A, Casjens R S and Hendrix R W 1998 *Science* **279** 1834
17. Diesiere F, Lucchini S and Brüssow H 1998 *Virology* **241** 345
18. Maurer P, Hohenester E and Engel J 1996 *Curr. Opinion Cell Biol.* **8** 609
19. Miyamoto H, Miyashita T, Okushima M, Nakano S, Morita T and Matsushiro A 1996 *Proc. Natl. Acad. Sci. USA* **93** 9657
20. Traub W, Yonath A and Segal D M 1969 *Nature (London)* **221** 914
21. Revel H R 1981 In *Bacteriophage assembly* (ed.) M S DuBow (New York: Liss) pp. 353–364
22. Earnshaw W C, Goldberg E B and Crowther R A 1979 *J. Mol. Biol.* **132** 101
23. Saitta B, Wang Y M, Renkart L, Zhang R Z, Pan T C, Timpl R and Chu M L 1991 *Genomics* **11** 145
24. Doolitle R F 1992 *Protein Sci.* **1** 191