Published in final edited form as:

Biomacromolecules. 2010 February 8; 11(2): 348. doi:10.1021/bm900894b.

Expanding the family of collagen proteins: Recombinant bacterial collagens of varying composition form triple-helices of similar stability

Chunying Xu, Zhuoxin Yu, Masayori Inouye, Barbara Brodsky, and Oleg Mirochnitchenko*

Department of Biochemistry, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ, 08854, USA

Abstract

The presence of the (Gly-Xaa-Yaa)_n open reading frames in different bacteria predicts the existence of an expanded family of collagen-like proteins. To further explore the triple-helix motif and stabilization mechanisms in the absence of hydroxyproline (Hyp), predicted novel collagen-like proteins from Gram-positive and -negative bacteria were expressed in *Escherichia coli* and characterized. Soluble proteins capable of successful folding and *in vitro* refolding were observed for collagen proteins from *Methylobacterium* sp 4-46, *Rhodopseudomonas palustris* and *Solibacter usitatus*. In contrast, all protein constructs from *Clostridium perfringens* were found predominantly in inclusion bodies. However, attachment of a heterologous N-terminal or C-terminal non-collagenous folding domain induced the *Clostridium perfringens* collagen domain to fold and become soluble. The soluble constructs from different bacteria had typical collagen triple-helical features and showed surprisingly similar thermal stabilities despite diverse amino acid compositions. These collagen-like proteins provide a resource for the development of biomaterials with new properties.

Introduction

Although collagen was thought to be a protein specific to multicellular animals, a bioinformatics search of available bacterial genome databases indicated the presence of about 100 open reading frames with typical collagen-like (Gly-Xaa-Yaa)_n sequences¹. This sequence pattern is suggestive of the collagen triple-helical structure in which close packing of the three supercoiled polypyrrolone-II-like polypeptide chains generates the constraint of Gly as every third residue^{2, 3}. Some of the bacterial collagen-like sequences have a high content of Pro, a feature also known to stabilize animal collagens by favoring the extended polypyrrolone II chain structure. But the putative collagen-like proteins differ significantly from animal collagens in one major respect: bacteria lack the prolyl hydroxylase which in animals serves to post-translationally hydroxylate Pro residues in the Yaa position to form Hydroxyproline (Hyp). In animal collagens, Hyp confers critical molecular stabilization due to stereoelectronic effects⁴ and its interaction with the hydration network^{5, 6}. A significant correlation is observed between the presence of the Hyp and thermal stability of collagens from different vertebrate and invertebrate organisms⁷. The presence of the (Gly-Xaa-Yaa)_n open reading frames in

*Author to whom correspondence should be addressed: Oleg Mirochnitchenko, Department of Biochemistry, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ, 08854, USA. Tel: 732-235-3469, Fax: 732-235-4783. mirochol@umdnj.edu.

Supporting Information The primary sequences of the bacterial proteins with collagenous domains (Figure S1) as well as amino acid compositions described in the manuscript (Table S1) are provided. Figure S2 contains Western blot of total protein lysate and inclusion bodies of *E. coli* cultures expressing *C. perfringens* collagen-like protein before and after trypsin treatment. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

bacteria suggests the production of many interesting proteins which offer potential for expanding our understanding of the triple-helix motif as well as facilitating large scale production of expressed collagen proteins for biotechnology applications.

In several pathogenic bacteria, the collagen-like proteins have been shown to be expressed and to form stable triple-helical proteins which play a role in pathogenicity. For example, Scl1 and Scl2 proteins from bacterium group A *Streptococcus pyogenes* (GAS) are expressed on the cell surface and Scl1 protein binding to $\alpha 2\beta 1$ integrin may mediate GAS internalization by human cell⁸. The Scl1 and Scl2 proteins form stable triple-helical structures when expressed as recombinant proteins⁹⁻¹², and an amino-terminal globular domain (V_{SP}) adjacent to the triple-helix domain appears to be important for efficient triple-helix assembly^{12, 13}. Other characterized prokaryotic collagen-like proteins include *Bacillus cereus* and *Bacillus anthracis* proteins associated with the exosporium with a probable role in spore–host interactions¹⁴⁻¹⁶; pneumococcal collagen-like protein A (PcIA) contributing to adhesion and invasion of host cells¹⁷; and a family of seven collagen-like proteins, called SclC-SclI from *Streptococcus equi* subspecies which are expressed upon infection of horses leading to the pathological condition known as strangles^{18, 19}.

To extend characterization of this unique group of proteins to non-pathogenic as well as pathogenic bacteria, several new predicted bacterial collagenous domains were expressed and investigated. Sequences were selected based on high probability to form a stable triple-helix structure. Non-collagenous sequences present on both sides of the (Gly-Xaa-Yaa)_n repeating sequence were included in the constructs, which were expressed using a high yield cold shock vector system in *E. coli*. Some of the proteins were soluble and correctly folded into stable triple-helical molecules. Other recombinant proteins formed inclusion bodies, but incorporation of their triple-helix domain together with a heterologous non-collagenous folding domain produced soluble chimeric molecules with proper folding and stable triple-helices. Surprisingly, the thermal stabilities of the triple-helix protein constructs from all bacteria were similar ($T_m=36.5-40^\circ\text{C}$), despite wide variations in amino acid composition, sequence, and length of the triple-helix domain, suggesting the potential of a variety of mechanisms for stabilization.

EXPERIMENTAL SECTION

Analysis of bacterial genomes

The NCBI microbial genome database was searched for annotated known and predicted collagen-like proteins with a relatively long (Gly-Xaa-Yaa)_n domain containing at least 35 repeats and lacking repetitive stretches of a single amino acid motif. For further selection, thermal stability of the collagen-like domains of selected proteins were predicted using the collagen stability calculator (jupiter.umdj.edu/collagen_calculator/)²⁰ and sequences showing regions of very low stability were excluded. ProtParam tool was used for evaluation of the AA content and physicochemical parameters.

DNA Amplification and Cloning

Full length genes for the collagen-like proteins were amplified using genomic DNA and corresponding primers (Table 1). Genomic DNA of *Clostridium perfringens* (*C. perfringens*), SM101 was provided by Hirofumi Nariya (Kagawa University, Japan), DNA of *Solibacter usitatus* (*S. usitatus*) was a gift from Cheryl R. Kuske (Los Alamos National Laboratory, NM), DNA of *Methylobacteria sp 4-46* (*M. sp 4-46*) was a gift from Christopher Marx (Harvard University, MA), DNA of *Rhodospirillum rubrum* (*R. rubrum*) and *Corynebacterium diphtheria* (*C. diphtheria*) were purchased from ATCC. Amplification

conditions were optimized for each pair of primers. Genes were cloned in pCR 2.1-TOPO vector and verified by DNA sequencing.

Each collagen domain CL, is denoted by the bacterial source, e.g. CL_{Rp} is the (Gly-Xaa-Yaa)_n region from *R. palustris*. The non-collagenous domains are denoted as (N)V for the N-terminal non-collagenous domain. For example, (N)V_{Sp} is N-terminal non-collagenous V domain from *S. pyogenes*. For obtaining of the full length *C. perfringens* construct (N)V_{Cp}-CL_{Cp}-(C)V_{Cp}, and partial *C. perfringens* constructs (N)V_{Cp}-CL_{Cp}(AA₁₋₂₄₂), CL_{Cp}(AA₅₄₋₂₄₂) protein and CL_{Cp}-(C)V_{Cp} (AA₅₄₋₄₀₃), corresponding fragments were amplified by PCR and re-cloned into the *E. coli* expression pColdII vector²¹ via NdeI/BamHI sites (the numbers in parenthesis correspond to the amino acids of the full size protein and are presented for the fragments the first time they are used). To obtain (N)V_{Sp}(AA₁₋₇₄)-CL_{Cp} and CL_{Cp}-(C)V_{Rp}(AA₁₂₇₋₂₁₂) recombinant proteins, fragments after amplification and assembly were cloned into the pColdIII²¹ via ApaI/BamHI and Nde/BamHI sites, correspondingly. To construct full length *S. usitatus* constructs (N)V_{Su}-CL_{Su}-(C)V_{Su}, and partial constructs (N)V_{Su}(AA₁₋₇₄)-CL_{Su}(AA₁₋₂₈₈), CL_{Su}(AA₄₃₋₂₈₈), CL_{Su}-(C)V_{Su}(AA₃₀₋₄₃₄), amplified fragments were cloned into the pColdII vector via NdeI/Hind III sites. pCold III vector already containing N-terminal domain of Scl2 *S. pyogenes* and SmaI/ApaI sites was used for obtaining (N)V_{Sp}(AA₁₋₇₄)-CL_{Su}(AA₄₃₋₂₈₈) protein. Full length (N)V_{Ms}-CL_{Ms}-(C)V_{Ms}, (N)V_{Ms}-CL_{Ms}(AA₂₃₋₂₇₁), CL_{Ms}(AA₂₃₋₂₇₁) *M. sp 4-46* recombinant proteins were constructed by amplification of the fragments and inserting them into pCold II vector using NdeI/BamHI sites. Full length and CL_{Rp}(AA₁₀₋₁₂₆)-(C)V_{Rp}(AA₁₂₇₋₂₁₂) *R. palustris* recombinant proteins were constructed by assembly of the corresponding fragments in pCold II vector using NdeI/BamHI sites. Fragments for the full length and N-terminal domain of Scl2 (N)V_{Sp}(AA₁₋₇₄)-CL_{Cd}(AA₁₋₂₂₂) fusion protein from *C. diphtheria* were cloned into pCold II and pCold III vectors already containing N-terminal domain via NdeI/BamHI and ApaI/BamHI sites, correspondingly. All recombinant proteins contained N-terminal His-Tag sequence for purification by affinity chromatography.

Protein expression and purification

Recombinant proteins were expressed in *E. coli* BL21 strain. For small scale purification and fractionation studies, bacterial cultures were grown in 10 ml of M9-casamino acid medium at 37°C until A₆₀₀ reached 0.9-1.0, then expression of the proteins was induced by 1mM isopropyl-D-thiogalactopyranoside and cultures were incubated on a shaker overnight at 20°C. To test protein solubility, cultures were centrifuged and pellets dissolved in 20 mM Na-phosphate, pH 7.4, 500 mM NaCl, sonicated and centrifuged. Supernatant was considered as a soluble fraction, whereas pellet undergoes additional cycle of resuspension, sonication and centrifugation. The final pellet was called the insoluble fraction. Proteins were analyzed by 12% SDS-PAGE.

Proteins were purified from one liter cultures, grown and induced as described for small scale production. Overnight cultures were centrifuged and resuspended in 20 mM Na-Phosphate, pH 7.4, 500 mM NaCl buffer with 10 mM β-mercaptoethanol, sonicated by 4-5×one minute bursts in Ultrasonic processor XL sonicator (Misonix). Extracts were centrifuged for 10 min at 12,000g, and after additional extraction of the pellets, supernatants were combined and centrifuged one hour at 45,000 rpm (rotor 50Ti, Beckman L7-55). Twenty five mM of imidazole was added to the extracts and it was loaded on 12.5 ml Ni-NTA agarose (QIAGEN) column. Column was washed sequentially with 50 ml of the binding buffer (Na-phosphate saline with 25 mM imidazole and 10 mM β-mercaptoethanol), 120 ml of the buffer with 58 mM of the imidazole and 50 ml of the buffer with 96 mM imidazole. Proteins were eluted with 30 ml of buffer containing 400 mM of imidazole. Proteins were dialyzed against Na-phosphate buffer, pH 8.6 with 50 mM glycine. Protein purity was checked by SDS-PAGE and MALDI-

TOF mass spectrometry. The following final yields from one liter of liquid culture were obtained for purified soluble proteins: (N)V_{Sp}-CL_{Cp} – 22.5 mg/l; CL_{Cp}-(C)V_{Rp} – 6.3 mg/l; CL_{Cp} – 9.8 mg/l; (N)V_{Ms}-CL_{Ms} – 22.3 mg/l; CL_{Ms} – 4.3 mg/l; (N)V_{Sp}-CL_{Su} – 15.3 mg/l; CL_{Su} – 4.3 mg/l; CL_{Rp}-(C)V_{Rp} – 4.3 mg/l; CL_{Rp} – 1 mg/l. We optimized further the growth conditions for two proteins, (N)V_{Sp}-CL_{Cp} and (N)V_{Sp}-CL_{Su} and corresponding CL domains. The use of rich 2×LB medium and induction of protein expression at A₆₀₀ reaching 5–6 lead to increased yields for CL_{Cp} and CL_{Su} up to 30–40 mg/l.

Trypsin Digestion

To test the recombinant proteins for sensitivity to trypsin digestion and to isolate the collagenous fragments, proteins were digested at room temperature with trypsin at a ratio 1:1000 (protein:enzyme) for different periods of time, and efficiency of the digestion was checked by electrophoresis. The reaction was terminated by addition of PMSF, followed by centrifugation and the supernatants were loaded onto a Superdex™ 200 gel filtration column (GE Healthcare). The purity of the fractions was checked by mass spectrometry.

Circular Dichroism Spectroscopy

AVIV Model 62DS spectropolarimeter (Aviv Associates Inc., Lakewood, NJ) was used for the recording of the CD spectra. Proteins were equilibrated in 1 mm cuvettes for at least 24 hrs at 4°C. Each scan was repeated three times and CD spectra were recorded from 195 to 260 nm with an average time of 5 s at 5 nm interval. Protein melting was monitored at 220 nm by increasing temperature in 0.33°C increments from 0 to 70°C. Proteins were maintained for 2 min at each temperature point and the average rate of temperature increase was 0.1°C/min. T_m is defined as the temperature at which the fraction folded is 50% in the curve fitted to the thermal transition. Proteins were refolded by decreasing the temperature from 70 to 0°C and process was recorded at 220 nm. The percentage of the refolding was determined as the ratio of the CD signal regained at 0°C after refolding to the initial signal before melting.

RESULTS

Selection of collagen-like protein candidates

To choose candidate proteins with collagen-like domains, a search was initiated in bacterial genomic databases of pathogenic and non-pathogenic bacteria for relatively long (Gly-Xaa-Yaa)_n domains (n>35), since the stability of the collagen triple helix is known to depend upon length of the protein up to a certain size after which T_m is length independent²⁰. Candidates were further analyzed in terms of predicted thermal stability of their CL domains using the collagen stability calculator (jupiter.umdj.edu/collagen_calculator/)²⁰, eliminating proteins predicted to have regions of low relative stability. The final protein candidates were grouped according to amino acid composition, and representative candidates with high percentage of charged residues, prolines, or predominantly polar residues were chosen for experimental investigations. It was also desirable to include proteins from non-pathogenic bacteria since none had been studied previously. Using these criteria, a set of putative proteins with collagen domains were selected from five bacteria (Table 1, Supplement 1): *C. perfringens*, a pathogenic gram-positive bacteria which is the causative agent of gas gangrene; *C. diphtheria*, a pathogenic rod-shaped gram-positive actinobacteria responsible for diphtheria; non-pathogenic gram-negative *M. sp 4-46* found mostly in soils or in plants which can utilize methanol emitted by the plants and stimulate plant development; non-pathogenic Gram-negative Acidobacteria (*S. usitatus*) which are abundant in soils; and non-pathogenic gram-negative Rhodopseudomonas (*R. palustris*), a phototrophic organism inhabiting marine environments and soil.

The collagen domains in these 5 bacteria contain distinctive amino acid compositions, with widely varying percentages of Pro, hydrophobic, charged, and polar residues (Fig. 1,

Supplement 2). All of the proteins have a high proportion of Pro, varying from 19.5-40% of all residues in the Xaa and Yaa positions. It is interesting to note that the proteins with the lower Pro contents have very high contents of acidic residues. The CL domain from *C. perfringens* (CL_{Cp}) has the lowest charge content (17% of all residues in the Xaa and Yaa positions) and a high content of Gln residues (29%) exclusively in the Yaa position, with a total of 37% polar residues, and 31% Pro. This contrasts with 35% charged residues, 37% Pro and 15% polar residues in the CL domain of *R. palustris*. Two proteins selected for the expression have acidic pI values (4.7 for *C. perfringens* and 5.4 for *S. usitatus*) whereas the rest of them are highly basic (pI 8.6, 9.3 and 8.85 for *M. sp 4-46*, *R. palustris* and *C. diphtheria*, respectively).

Repeating sequence patterns are notable in most of these proteins. For example, CL_{Cp} has 6 full and 3 partial repeats of the Gln rich sequence GP[RQ]GP[VIR]G[PL]QGEQGPQGERGF while eight full repeating charge sequences of the form GPKGEP are present in CL_{Ms}. The CL region from *S. usitatus* (CL_{Su}) has 2 large Ala-rich repeats at the N-terminus GPAGPAGPQGPAGP as well as numerous imperfect repeats (see Supporting information, Fig. S1). Other repeating sequences are seen in the CL domains of *C. diphtheria* (CL_{Cd}) and *R. palustris* (CL_{Rp}). This contrasts with the absence of repeating sequences in animal collagens, although there are periodicities of charged and hydrophobic residues²². There are different numbers of these repeats in different strains²³, and it is not clear at this stage if any of the bacterial collagen repeats are important for the function and stability of the proteins or if they are related to the evolution of these collagen-like sequences.

These proteins from 5 different bacteria were selected on the basis of the characteristics of their (Gly-Xaa-Yaa)_n domains. But all animal collagens and the few bacterial collagen-like proteins characterized so far have non-triple-helix regions surrounding the triple-helix domain which are necessary for trimerization, nucleation and registration of the triple-helix^{13, 24}. The full length protein in the open reading frame containing the (Gly-Xaa-Yaa)_n sequence was included for expression in all cases (Table 1).

Expression of bacterial collagen-like proteins in *E. coli*

Genes for these candidate proteins (Table 1) were cloned from genomic DNA and expressed in *E. coli* BL21 strain using cold-shock expression vectors. Initial expression plasmids were constructed using full length proteins including N and C-terminal non-collagenous domains (Fig. 2). Predicted signal peptide coding regions were not included in the constructs when they could be identified. No inducible expression was observed for the *C. diphtheria* recombinants and this bacterial protein was not further characterized (Fig. 2). The problematic expression could be due to a 9 residue interruption in the (Gly-Xaa-Yaa)_n sequence of the CL_{Cd} region, since such breaks in the repeating triplet pattern have been shown to lead to the disturbance and destabilization of the triple helix²⁵ and possibly to the protein instability. Good expression was observed for the recombinant collagen-like proteins of the other four bacteria, and their purification and characterization were carried out.

Cell fractionation indicated that the full-length proteins from three bacteria (*M. sp 4-46*, *S. usitatus* and *R. palustris*) were present in the soluble fraction (Fig. 2 and 3). All recombinant protein constructs from *M. sp 4-46* were soluble including the collagen domain alone. The recombinant protein from *R. palustris* was almost fully soluble as a full length construct or as CL_{Rp} with the C-terminus (C)V_{Rp}, whereas protein from *S. usitatus* become more soluble after the deletion of the C-terminal domain. Partial solubility was seen for the other recombinant proteins constructs of *S. usitatus*. The protein from the *C. perfringens* was found in inclusion bodies (Fig. 2). Neither the CL_{Cp} domain alone nor CL_{Cp} with one or both of its terminal domains, were soluble.

Formation of chimeric proteins to promote folding

Since insoluble proteins in inclusion bodies have been linked to misfolding²⁶, the insoluble CL_{Cp} collagen domain was fused with two potential folding domains from the collagen proteins of other bacteria. In one construct, an additional C-terminal domain of *R. palustris*, (C)V_{Rp} was attached to the CL_{Cp} domain, in another, it was fused with a preceding N-terminal globular (N)V_{Sp} domain of the Scl2 collagen-like protein from *S. pyogenes* (Fig. 2). Constructs containing CL_{Cp} domain were now expressed in the soluble fraction, suggesting these additional chimeric domains were effective in facilitation of the correct assembly of the collagen-like domain from *C. perfringens*. There appears to be specificity in the location of the folding domains since the V domain of *R. palustris* promoted folding when on the C-terminal but not the N-terminal side of CL_{Cp}. Attachment of the (N)V_{Sp} domain to the CL_{Su} domain of *S. usitatus* also led to efficient production in the soluble fraction, suggesting proper protein assembly. No *C. diphtheria* protein induction was observed even after fusion of its CL domain with (N)V_{Sp}. Analysis of the protein samples by SDS PAGE is shown in Fig. 3.

Trypsin resistance of the collagenous domains

The collagen triple-helix confers resistance to digestion by trypsin as well as by other proteinases²⁷, and such enzyme digestion can be used to verify the presence of the triple-helix as well as to purify collagen domains. To probe the conformation of the expressed bacterial collagen constructs, recombinant proteins were purified from soluble fractions by affinity chromatography on Ni-NTA agarose and then digested with trypsin for different period of times at room temperature. The trypsin-digested products had mobilities expected for the isolated collagen triple-helix domains of the 4 bacterial proteins (Fig. 3), except for the mobility of the CL domain from *R. palustris* which was slightly slower than predicted. Mass spectroscopy confirmed the purity and molecular weight of the (Gly-Xaa-Yaa)_n collagen domains of the proteins: 18,870.19 Da for CL_{Cp}; 21,667.53 Da for CL_{Su}; and 14,967.28 Da for CL_{Ms}. The size of the CL_{Rp} was 13,335.87 Da, larger than the 11,275.48 Da theoretical value of the collagen domain alone, due to cleavage at an Arg site 24 amino acids into the N-terminal domain. Insertion of an Arg at the end of the CL domain is planned to obtain only the triple-helical fragment.

These results strongly suggest that soluble collagen-like proteins from four bacteria expressed in *E. coli* contain CL domains which are in a trypsin resistant triple-helical conformation. The conformation of the CL domains in non-soluble protein fractions was also approached using trypsin digestion. As it is shown in Fig. S2 (Supporting information), a significant portion of the inclusion bodies of the (N)V_{Cp}-CL_{Cp} has been digested by trypsin and no protease-resistant fragment corresponding by SDS-PAGE mobility to the CL domain was observed. These data support the conclusion that non-soluble recombinant proteins with collagen-like domains are likely to represent aggregates of misfolded protein.

Circular Dichroism Spectroscopy: Conformation, Thermal Stability, and Refolding

CD spectra of purified CL_{Su}, CL_{Rp} and CL_{Cp} collagen domains obtained by trypsin digestion, as well as CL_{Ms} expressed as only the collagenous domain gave typical collagen-like features at 4°C, with a maximum at 220 nm and a minimum near 198 nm (Fig. 4, insets). It is difficult to obtain precise concentrations and mean residue ellipticity value for collagen domains which have no aromatic residues. The Rpn value (ratio of positive to negative peak) was used to estimate their triple-helix content²⁸. Values suggesting a fully triple-helical molecule (Rpn~0.11)²⁸ were seen for the CL_{Cp} and CL_{Su}. The lower Rpn value recorded for CL_{Ms} and the negative value of the 220 nm peak for CL_{Rp} suggest some perturbation to the triple helix or partial degradation (Table 2).

The CD spectra of proteins from single bacterial species, *M. sp 4-46* (N)V_{MS}-CL_{MS} and *R. palustris* CL_{RP}-(C)V_{RP} as well as several chimeric proteins containing a CL domain from one bacteria together with a folding domain from another bacteria, CL_{CP}-(C)V_{RP} and (N)V_{SP}-CL_{SU}, have been studied. CD spectroscopy of all chimeric proteins and (N)V_{MS}-CL_{MS} showed collagen-like features with a maximum at 220 nm and a minimum near 198 nm but with much lower magnitude than seen for isolated CL domains (Fig. 4 and 5, insets). The non-collagenous domains contribute to the spectrum and are likely cancel out some of the collagen-like signal¹². The *R. palustris* construct CL_{RP}-(C)V_{RP} shows only a shoulder at 220 nm (Fig. 5, inset), which is consistent with it having the shortest CL domain and the prediction of a helical coiled coil structure in the (C)V_{RP} domain (85 residues) (Fig. 4).

Thermal unfolding of recombinant proteins was followed by monitoring the CD signal at 220 nm with increasing temperature (pH 8.6). Very sharp thermal transitions were observed for CL domains of *S. usitatus* and *C. perfringens*. Broader transitions were detected for the CL domains of the *M. sp 4-46* and *R. palustris* (Fig. 4), which together with low Rpn values may suggest partial unfolding or heterogeneity of the material. T_m values were in 35 to 39°C range (Table 2). Slightly higher T_m values were observed when collagenous domains were covalently attached to the folding domains compared with CL domains alone, indicating a relatively small stabilizing effect of the non-collagenous domains (Fig. 5). The only exception was *M. sp 4-46*, for which CL domain alone has almost 5°C lower T_m than the same protein with its own N-terminal folding domain (35.0°C versus 40.3°C).

CD spectra and melting curves were also obtained at pH 2.2, to compare with corresponding measurements at pH 8.6. CD spectra were similar at different pH values, but a significant decrease in thermal stability was observed at low pH for the CL (T_m=38.5 to 27°C), CL_{MS} (T_m=40.3 to 28.3°C) and CL_{RP} (T_m=37 to 32°C) (Fig. 4, Table 2). It is also interesting to note that melting curves of the CL_{MS} and CL_{RP} were sharper upon melting at pH 2.2 than at pH 8.6. The three CL proteins which show a strong dependence of the stability from pH have a very high proportion of charged residues, constituting 34%, 34% and 35% of all Xaa, Yaa residues in (Gly-Xaa-Yaa) for CL_{MS}, CL_{SU} and CL_{RP}, respectively. In contrast to the high pH dependent decrease in T_m for the CL domains from the other three bacteria, only a slight decrease in the T_m value (1.6°C) was observed for CL from *C. perfringens*, which has the lowest charge content. Among three CL domains with high percentage of the charged residues, CL_{MS} and CL_{SU} have almost equal quantity of negatively and positively charged residues, whereas CL_{RP} has 2.4 times more positively charged than negatively charged residues (see Table S1, Supporting information).

The ability of purified constructs to refold *in vitro* was investigated by monitoring the CD signal at 220 nm upon cooling of the samples from 70°C to 0°C at the same rate as heating (~0.1°C/min) (Fig. 5). CL domains isolated from all bacteria by trypsin digestion or by expression of the CL domain alone (CL_{MS}) were not able to refold (data not shown). Most of the constructs with non-triple-helical domains adjacent to the CL domain showed recovery of some of the CD signal (Fig. 5A-D). The efficiency of refolding varied among the different constructs, with complete refolding in the case of the CL_{RP}-V_{RP} and minimal refolding in the case CL_{CP}-V_{RP} (Fig. 5). It is interesting that (C)V_{RP} domain was extremely effective in refolding its own CL domain but not in the refolding of the heterogeneous CL from *C. perfringens*.

DISCUSSION

The presence of more than 100 bacterial protein sequences containing (Gly-Xaa-Yaa)_n domains in genomic databases raises the possibility of a whole new family of collagen-like proteins which could expand our understanding of this motif and provide potential sources of collagen-based biomaterials. Bacterial collagen-like proteins from *S. pyogenes* and *B.*

anthrax have been well characterized⁹⁻¹⁶, and these are extended here to include new proteins from non-pathogenic as well as pathogenic bacteria. Previously, large amounts of the Scl2 collagen like protein from *S. pyogenes* were obtained using expression in a cold shock vector at room temperature^{12, 29}, and this method was successfully applied here for expression of novel proteins with (Gly-Xaa-Yaa)_n domains from four different bacteria.

The four CL domains all have in common the presence of Gly as every third residues, but they differ dramatically in their amino acid composition, sequence, and length of the (Gly-Xaa-Yaa)_n domain. Despite the diverse sequences and the absence of Hyp in these bacterial collagen proteins, all of the (Gly-Xaa-Yaa)_n CL domains show characteristics of a classic triple helix structure, including resistance to trypsin digestion and the typical collagen triple-helix CD spectrum. It was striking to observe that all 4 collagen-like domains with varying lengths and compositions have very high thermal stabilities, ranging from 35-39°C, similar to the values of various *S. pyogenes* Scl2 and Scl1 proteins, the *B. anthrax* Bcl1 protein, and mammalian collagens. The proteins studied here were preselected in part based upon the prediction of a stable triple-helix CL domain^{30, 31}, but the similarity of all these thermal stabilities is still unexpected.

Examination of the sequences of the collagen domains from these 4 bacteria suggests features of the mechanisms of triple-helix structure stabilization. A high Pro content must be a major contributor to stability. All of the proteins have a high proportion of Pro, varying from 19.5-40% of all residues in the Xaa and Yaa positions. Three of the CL proteins are rich in charged residues and the significant decrease in T_m value for the three charge-rich proteins at acidic conditions confirms the importance of electrostatic interactions in these three proteins. The collagen domain from *C. perfringens* has the lowest charge content (17%) and absence of any significant effect of the low pH on thermal stability of this protein supports the notion that electrostatic interactions do not play a significant role in stabilization of *C. perfringens* CL domain.

Animal collagens or proteins with collagen-like domains (e.g. collectins, macrophage scavenger receptor) have been shown in many cases to have a non-collagenous trimerization domain adjacent to the (Gly-Xaa-Yaa)_n sequence^{24, 32-34}. Studies on bacterial collagen-like proteins suggest they may follow similar principles, but that factors specific to each bacteria may play a role. Successful folding and refolding of collagen domains with their adjacent non-collagenous regions were previously reported for *S. pyogenes* Scl1 and Scl2 proteins¹³, and similar results were seen for the collagen proteins from *M. sp 4-46* and *R. palustris*, which were soluble as full length constructs and could refold *in vitro*. The full length proteins from the other two bacteria, *C. perfringens* and *S. usitatus*, both containing N- and C-terminal non-collagenous sequences which might be expected to act as trimerization/folding domains, but, surprisingly, the proteins were insoluble and found in inclusion bodies.

Bacterial inclusion bodies are refractile aggregates of protease-resistant misfolded protein that often occur in recombinant bacteria upon overexpression of cloned genes^{35, 36}. Aggregation of recombinant proteins is most likely due to a limiting amount of chaperones leading to the intermolecular association of exposed hydrophobic surfaces before the protein folding can be completed³⁷⁻³⁹. Treatment of insoluble *C. perfringens* aggregates with trypsin did not indicate the presence of a resistant CL domain suggesting that insoluble CL domains are not correctly folded into triple-helical structures. It is not clear why the native non-collagenous domains are not effective for these two bacteria, but it is possible that folding of these collagen-like proteins in *E. coli* requires specific chaperones or interactive proteins which assist the protein folding in the native host. However, a construct including either an N-terminal *S. pyogenes* V_{Sp} folding domain or a C-terminus *R. palustris* V_{Rp} folding domain attached to CL_{Cp} or CL_{Su} was sufficient to ensure folding, solubility, and *in vitro* refolding. The ability of the heterologous

non-collagenous domain to promote folding and solubility supports the argument that solubility is dependent on correct folding.

Recent studies indicate that high yield bacterial production of nontraditional sources of proteins with properties similar to human collagen might provide an alternative technology for biomaterial applications. Here it was demonstrated that predicted stable CL-domains from several non-pathogenic and pathogenic bacterial species can be efficiently expressed and form stable triple-helices in cold-shock expression system if they fold correctly or they can be induced to do so upon attachment of effective folding domains. Large scale production of these proteins in industrial bioreactors avoiding oxygen and nutrient limitations, and capable of reaching high *E. coli* densities will allow increasing production of the recombinant bacterial collagens up to several grams per liter of liquid culture. The availability of the collagen-like proteins with different amino acid composition capable of forming stable triple-helical structures provides a foundation for the development of new biomaterials.

CONCLUSIONS

In summary, this manuscript reports cloning, high yield expression and preliminary characterization of four new collagen-like proteins from pathogenic and non-pathogenic bacterial species. Analysis of the primary amino acid sequence of these proteins indicates a wide variety of compositions and allows categorizing them in two groups, with low and higher percentage of charged residues. Charged residues in proteins from the second group are shown to contribute significantly to the protein structure stability. Another important feature of these proteins is their high Pro content and absence of secondary modifications. Using recombinant protein technology the protein domains required for expression of soluble CL domains were determined. Several CL domains require construction of the protein chimeras combining folding domain from one species and CL domain from another for this purpose. Biophysical characterization of the CL domains indicates that they are folded in stable triple-helix structures with similar melting temperatures in 35-39°C range. These bacterial collagen-like proteins offer an opportunity to create stable triple-helix protein products for biomaterial applications in a high yield bacterial expression system. The bacterial approach has several additional advantages, such as easy genetic manipulation of the primary sequence, leading to the modification of biological properties⁸ and the formation of high-order structures by these recombinant molecules⁴⁰. Further investigation is required to study properties of these proteins in terms of suitability for specific biomaterial applications as well as to characterize their biological effects.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Bethany Walton and Eileen Hwang for the help in sequence analysis and obtaining CD spectra and John A.M. Ramshaw for help in determining candidate proteins for our analysis. This work was supported by NIH grant R21 EB007198 (B.B.).

REFERENCES

1. Rasmussen M, Jacobsson M, Bjorck L. Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins. *J Biol Chem* 2003;278(34):32313–6. [PubMed: 12788919]
2. Ramachandran GN, Kartha G. Structure of collagen. *Nature* 1955;176(4482):593–5. [PubMed: 13265783]

3. Rich A, Crick FH. The molecular structure of collagen. *J Mol Biol* 1961;3:483–506. [PubMed: 14491907]
4. Jenkins CL, Raines RT. Insights on the conformational stability of collagen. *Nat Prod Rep* 2002;19(1):49–59. [PubMed: 11902439]
5. Bella J, Eaton M, Brodsky B, Berman HM. Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science* 1994;266(5182):75–81. [PubMed: 7695699]
6. Privalov PL. Stability of proteins. Proteins which do not present a single cooperative system. *Adv Protein Chem* 1982;35:1–104. [PubMed: 6762066]
7. Burjanadze TV, Kisiriya EL. Dependence of thermal stability on the number of hydrogen bonds in water-bridged collagen structure. *Biopolymers* 1982;21(9):1695–701. [PubMed: 7126753]
8. Caswell CC, Barczyk M, Keene DR, Lukomska E, Gullberg DE, Lukomski S. Identification of the first prokaryotic collagen sequence motif that mediates binding to human collagen receptors, integrins $\alpha 2\beta 1$ and $\alpha 1\beta 1$. *J Biol Chem* 2008;283(52):36168–75. [PubMed: 18990704]
9. Lukomski S, Nakashima K, Abdi I, Cipriano VJ, Ireland RM, Reid SD, Adams GG, Musser JM. Identification and characterization of the scl gene encoding a group A *Streptococcus* extracellular protein virulence factor with similarity to human collagen. *Infect Immun* 2000;68(12):6542–53. [PubMed: 11083763]
10. Rasmussen M, Eden A, Bjorck L. SclA, a novel collagen-like surface protein of *Streptococcus pyogenes*. *Infect Immun* 2000;68(11):6370–7. [PubMed: 11035747]
11. Rasmussen M, Bjorck L. Unique regulation of SclB - a novel collagen-like surface protein of *Streptococcus pyogenes*. *Mol Microbiol* 2001;40(6):1427–38. [PubMed: 11442840]
12. Mohs A, Silva T, Yoshida T, Amin R, Lukomski S, Inouye M, Brodsky B. Mechanism of stabilization of a bacterial collagen triple helix in the absence of hydroxyproline. *J Biol Chem* 2007;282(41):29757–65. [PubMed: 17693404]
13. Xu Y, Keene DR, Bujnicki JM, Hook M, Lukomski S. Streptococcal Scl1 and Scl2 proteins form collagen-like triple helices. *J Biol Chem* 2002;277(30):27312–8. [PubMed: 11976327]
14. Todd SJ, Moir AJ, Johnson MJ, Moir A. Genes of *Bacillus cereus* and *Bacillus anthracis* encoding proteins of the exosporium. *J Bacteriol* 2003;185(11):3373–8. [PubMed: 12754235]
15. Sylvestre P, Couture-Tosi E, Mock M. A collagen-like surface glycoprotein is a structural component of the *Bacillus anthracis* exosporium. *Mol Microbiol* 2002;45(1):169–78. [PubMed: 12100557]
16. Steichen C, Chen P, Kearney JF, Turnbough CL Jr. Identification of the immunodominant protein and other proteins of the *Bacillus anthracis* exosporium. *J Bacteriol* 2003;185(6):1903–10. [PubMed: 12618454]
17. Paterson GK, Nieminen L, Jefferies JM, Mitchell TJ. PclA, a pneumococcal collagen-like protein with selected strain distribution, contributes to adherence and invasion of host cells. *FEMS Microbiol Lett* 2008;285(2):170–6. [PubMed: 18557785]
18. Karlstrom A, Jacobsson K, Flock M, Flock JI, Guss B. Identification of a novel collagen-like protein, SclC, in *Streptococcus equi* using signal sequence phage display. *Vet Microbiol* 2004;104(34):179–88. [PubMed: 15564026]
19. Karlstrom A, Jacobsson K, Guss B. SclC is a member of a novel family of collagen-like proteins in *Streptococcus equi* subspecies *equi* that are recognised by antibodies against SclC. *Vet Microbiol* 2006;114(12):72–81. [PubMed: 16384667]
20. Persikov AV, Ramshaw JA, Brodsky B. Prediction of collagen stability from amino acid sequence. *J Biol Chem* 2005;280(19):19343–9. [PubMed: 15753081]
21. Qing G, Ma LC, Khorchid A, Swapna GV, Mal TK, Takayama MM, Xia B, Phadtare S, Ke H, Acton T, Montelione GT, Ikura M, Inouye M. Cold-shock induced high-yield protein production in *Escherichia coli*. *Nat Biotechnol* 2004;22(7):877–82. [PubMed: 15195104]
22. Hulmes DJ, Miller A, Parry DA, Piez KA, Woodhead-Galloway J. Analysis of the primary structure of collagen for the origins of molecular packing. *J Mol Biol* 1973;79(1):137–48. [PubMed: 4745843]
23. Han R, Zwiefka A, Caswell CC, Xu Y, Keene DR, Lukomska E, Zhao Z, Hook M, Lukomski S. Assessment of prokaryotic collagen-like sequences derived from streptococcal Scl1 and Scl2 proteins as a source of recombinant GXY polymers. *Appl Microbiol Biotechnol* 2006;72(1):109–15. [PubMed: 16552563]

24. Khoshnoodi J, Cartiailler JP, Alvares K, Veis A, Hudson BG. Molecular recognition in the assembly of collagens: terminal noncollagenous domains are key recognition modules in the formation of triple helical protomers. *J Biol Chem* 2006;281(50):38117–21. [PubMed: 17082192]
25. Brodsky B, Thiagarajan G, Madhan B, Kar K. Triple-helical peptides: an approach to collagen conformation, stability, and self-association. *Biopolymers* 2008;89(5):345–53. [PubMed: 18275087]
26. de Groot NS, Espargaro A, Morell M, Ventura S. Studies on bacterial inclusion bodies. *Future Microbiol* 2008;3:423–35. [PubMed: 18651814]
27. Bruckner P, Prockop DJ. Proteolytic enzymes as probes for the triple-helical conformation of procollagen. *Anal Biochem* 1981;110(2):360–8. [PubMed: 7015914]
28. Feng Y, Melacini G, Taulane JP, Goodman M. Collagen-based structures containing the peptoid residue N-isobutylglycine (Nleu): synthesis and biophysical studies of Gly-Pro-Nleu sequences by circular dichroism, ultraviolet absorbance, and optical rotation. *Biopolymers* 1996;39(6):859–72. [PubMed: 8946805]
29. Yoshizumi A, Yu Z, Silva T, Thiagarajan G, Ramshaw JA, Inouye M, Brodsky B. Self-association of streptococcus pyogenes collagen-like constructs into higher order structures. *Protein Sci* 2009;18(6):1241–51. [PubMed: 19472339]
30. Shah NK, Ramshaw JA, Kirkpatrick A, Shah C, Brodsky B. A host-guest set of triple-helical peptides: stability of Gly-X-Y triplets containing common nonpolar residues. *Biochemistry* 1996;35(32):10262–8. [PubMed: 8756681]
31. Persikov AV, Ramshaw JA, Kirkpatrick A, Brodsky B. Amino acid propensities for the collagen triple-helix. *Biochemistry* 2000;39(48):14960–7. [PubMed: 11101312]
32. Pakkanen O, Hamalainen ER, Kivirikko KI, Myllyharju J. Assembly of stable human type I and III collagen molecules from hydroxylated recombinant chains in the yeast *Pichia pastoris*. Effect of an engineered C-terminal oligomerization domain foldon. *J Biol Chem* 2003;278(34):32478–83. [PubMed: 12805365]
33. Zhang P, McAlinden A, Li S, Schumacher T, Wang H, Hu S, Sandell L, Crouch E. The amino-terminal heptad repeats of the coiled-coil neck domain of pulmonary surfactant protein d are necessary for the assembly of trimeric subunits and dodecamers. *J Biol Chem* 2001;276(23):19862–70. [PubMed: 11279100]
34. Boudko SP, Engel J, Bachinger HP. Trimerization and triple helix stabilization of the collagen XIX NC2 domain. *J Biol Chem* 2008;283(49):34345–51. [PubMed: 18845531]
35. Prouty WF, Karnovsky MJ, Goldberg AL. Degradation of abnormal proteins in *Escherichia coli*. Formation of protein inclusions in cells exposed to amino acid analogs. *J Biol Chem* 1975;250(3):1112–22. [PubMed: 1089651]
36. Carrio MM, Villaverde A. Construction and deconstruction of bacterial inclusion bodies. *J Biotechnol* 2002;96(1):3–12. [PubMed: 12142138]
37. Rinas U, Bailey JE. Overexpression of bacterial hemoglobin causes incorporation of pre-beta-lactamase into cytoplasmic inclusion bodies. *Appl Environ Microbiol* 1993;59(2):561–6. [PubMed: 8382023]
38. Lorimer GH. A quantitative assessment of the role of the chaperonin proteins in protein folding in vivo. *FASEB J* 1996;10(1):5–9. [PubMed: 8566548]
39. King J, Haase-Pettingell C, Robinson AS, Speed M, Mitraki A. Thermolabile folding intermediates: inclusion body precursors and chaperonin substrates. *FASEB J* 1996;10(1):57–66. [PubMed: 8566549]
40. Yoshizumi A, Yu Z, Silva T, Thiagarajan G, Ramshaw J, Inouye M, Brodsky B. Self-association of *Streptococcus pyogenes* Collagen-like constructs into Higher Order Structures. *Protein Science*. 2009 in press.

Figure 1. Pie chart representation of the non-Gly residues composition of the bacterial collagen-like domains

Amino acids only at Xaa and Yaa positions have been considered. The following groups are shown: Polar (Ser, Thr, Cys, Asn, Gln), Charged (Asp, Glu, Lys, Arg, His), Small (Gly and Ala), Pro and Hydrophobic (Val, Ile, Leu, Met, Phe, Trp, Cys). For comparison, the amino acid composition of the α I (I) chain of human type I collagen is shown.

Figure 2. Schematic diagram of recombinant proteins with bacterial collagen-like domains, constructed for the expression in E coli

Numbers, length of each domain in AA; black filled boxes, CL domains of corresponding proteins; empty boxes, N- and C-terminal domains; vertical empty box, collagen-like sequence interruption; boxes with line dashed patterns, V domain from Scl2 of *S. pyogenes* (V_{Sp}) and C-terminal V domain from *R. palustris* (V_{Rp}). Also shown are the expression levels (Ex) and solubility (Sol), determined by fractionation and SDS PAGE.

Figure 3. Expression of recombinant proteins in *E. coli* and resistance to trypsin digestion
 Recombinant plasmids were transfected in BL21 strain. Protein expression was induced by 1 mM isopropyl-D-thiogalactopyranoside overnight at 20°C. Proteins were purified on an Ni-NTA agarose (QIAGEN) column and dialyzed against Na-phosphate buffer, pH 8.6 with 50 mM glycine. To test for sensitivity of the recombinant proteins to trypsin digestion, they were digested at room temperature with trypsin at ratio 1:1000 (protein:enzyme) for one hour (or different period of time where shown) and efficiency of the digestion was checked by electrophoresis. Arrows indicate trypsin resistant species.

Figure 4. Thermal stability of the recombinant bacterial collagen-like domains
 CD thermal transitions of the corresponding CL domains at pH 8.6 (filled squares, continuous lines) and pH 2.2 (empty circles, dotted lines) were monitored at 220 nm. CD spectra of proteins are shown as insets.

Figure 5. Thermal transitions of the recombinant proteins determined by the monitoring CD signal at 220 nm

The arrow indicates the direction of temperature change with $\rightarrow$, for the unfolding curve with increasing temperature and $\leftarrow$ for the refolding curve with decreasing temperature. The heating rates were $\sim 0.1^{\circ}\text{C/min}$ in both directions. CD spectra of each protein are shown as an inset.

Table 1

Predicted bacterial collagen-like proteins

Bacteria	Protein	MW (Kd) / PI	Primers, forward/reverse
<i>Clostridium perfringens</i> , SM101	ABG86771.1	42.1/4.7	5'AGAAGCTCCAATGGCAAAGGAAGATGA3'/ 5'ACTCATTCAACTGGAGGCGTATGCATTTC'
<i>Solibacter usitatus</i>	YP_822627.1	40.8/5.4	5'TCCCGATTGAGGCGAAGCAAA CTT3'/ 5'TACGCGATGACGCAT TGA GGGAAA3'
<i>Methylobacterium</i> , <i>sp</i> 4-46	ACA18713.1	33.5/8.6	5'AATCTCGACCGCAAGGACCTCTAC3'/ 5'ACATCCGCAAGGCGAAGCAAT3'
<i>Rhodopseudomonas palustris</i>	YP_0019930	22.1/9.3	5'AAT TGA AGC CGT CAC GCA AGC TCT3'/ 5'TGACGGAACATCAAGACGCTGTCAA3'
<i>Corynebacterium diphtheriae</i>	CAE50366.1	25.8/8.85	5' AAC TTT CCC GCC GTG TTG TCC AAT3'/ 5'TGC AAG AAT TGT TGG GCC ATG CGA3'

Table 2

Properties of the bacterial collagen-like recombinant proteins expressed in *E. coli*.

Bacteria	Recombinant Protein ^a	Trypsin Resistance ^b	T _m , °C	CD, Rpn ^c
<i>C. perfringens</i>	V _{Sp} -CL _{Cp} CL _{Cp} -V _{Rp} CL _{Cp} ^d CL _{Cp} , pH 2.2	+ + + Nt ^e	39.6 40.2 38.8 37.2	0.13
<i>S. usitatus</i>	V _{Sp} -CL _{Su} CL _{Su} ^d CL _{Su} , pH 2.2 ^d	+ + Nt	39.4 38.5 27.0	0.11
<i>M. sp 4-46</i>	V _{Ms} -CL _{Ms} CL _{Ms} CL _{Ms} , pH 2.2	+ + Nt	40.3 35.0 28.3	0.06
<i>R. palustris</i>	CL _{Rp} -V _{Rp} CL _{Rp} ^d CL _{Rp} , pH 2.2 ^d	+ + Nt	37.5 37.0 32.0	N/a ^g
<i>S. pyogenes</i> ^f Scl2	V _{Sp} -CL _{Sp} CL _{Sp} ^d CL _{Sp} , pH 2.2 ^d	+ + Nt	35.6 35.9 25.7	0.11

^a - V_{Sp} N-terminal globular domain from *S. pyogenes* Scl2 protein; V, N-terminal domain from *M. sp 4-46* protein; V_{Rp}, C-terminal domain from *R. palustris* protein; CL, collagen-like domain

^b - CL domain of the recombinant V-CL or CL-V proteins were resistant to the trypsin digestion

^c - Rpn is the ratio of the positive at 220 nm to the negative peak at 198 nm

^d - CL domains were purified by trypsin digestion

^e - not tested

^f - data for *S. pyogenes* Scl2 protein and its CL domain presented for comparison ¹²

^g - not available due to the negative value of the peak at 220 nm