Collagens are extended trimeric proteins composed of the repetitive sequence glycine-X-Y. A collagen-related structural motif (CSM) containing glycine-X-Y repeats is also found in numerous proteins often referred to as collagen-like proteins. Little is known about CSMs in bacteria and viruses, but the occurrence of such motifs has recently been demonstrated. Moreover, bacterial CSMs form collagen-like trimers, even though these organisms cannot synthesize hydroxyproline, a critical residue for the stability of the collagen triple helix. Here we present 100 novel proteins of bacteria and viruses (including bacteriophages) containing CSMs identified by in silico analyses of genomic sequences. These CSMs differ significantly from human collagens in amino acid content and distribution; bacterial and viral CSMs have a lower proline content and a preference for proline in the X position of GXY triplets. Moreover, the CSMs identified contained more threonine than collagen, and in 17 of 53 bacterial CSMs threonine was the dominating amino acid in the Y position. Molecular modeling suggests that threonines in the Y position make direct hydrogen bonds to neighboring backbone carbonyls and thus substitute for hydroxyproline in the stabilization of the collagen-like triple-helix of bacterial CSMs. The majority of the remaining CSMs were either rich in proline or rich in charged residues. The bacterial proteins containing a CSM that could be functionally annotated were either surface structures or spore components, whereas the viral proteins generally could be annotated as structural components of the viral particle. The limited occurrence of CSMs in eubacteria and lower eukaryotes and the absence of CSMs in archaeabacteria suggests that DNA encoding CSMs has been transferred horizontally, possibly from multicellular organisms to bacteria.

However, non-fibrillar collagens as well as proteins with shorter collagen-like regions also exist, and a collagen-related structural motif (CSM) has been identified in proteins of bacteria, bacteriophages, and viruses (2–5). Bacteria and viruses are generally believed to be unable to synthesize hydroxyproline, a residue regarded as essential for the stabilization of a triple-helical structure. However, the absolute requirement of hydroxyproline for triple-helix formation has lately been challenged (6, 7), and alternative means for the stabilization of triple-helical collagen have been proposed (8, 9). Importantly, it has been shown that three different bacterial CSMs trimerize, despite the lack of hydroxyprolines (2, 10).

This study was undertaken to investigate the occurrence of proteins containing CSMs in bacteria and viruses through an in silico approach. We found that CSMs are encoded by a minority of bacteria and bacteriophages and that these CSMs differ from human collagens in several important aspects. A novel mechanism for the stabilization of bacterial CSMs is presented. We also propose that horizontal gene transfer has contributed to the evolution of CSMs.

EXPERIMENTAL PROCEDURES

The CSMs were identified by downloading 56 bacterial proteomes from The Institute of Genomic Research (TIGR.org) homepage using the batch download function and pattern searches in-house, using MacVector (version 6.5.3; Oxford Molecular Ltd.). The pattern used was (GPP)7 with 12 allowed mismatches. Matches containing Gly in the Pro positions were excluded. The CSMs were then used in unfiltered tBLASTn searches from the NCBI microbial genome homepage against 136 eubacterial genomes, 15 archeabacterial genomes, 30 genomes of lower eukaryotes, and against available viral genomes. Obtained hits were analyzed with the pattern in-house. Open reading frames flanking the bacterial CSM-encoding genes were analyzed by BLASTp to determine whether the gene was phage-encoded. Homology models of proline-rich CSMs were constructed on a trimeric (GPP)n template from a 1.3-Å x-ray structure (11). The models were built by global energy minimization using the template as constraint. After the model structures had been minimized, conformations of all threonine side chains were sampled through an exhaustive systematic search of the χ-angles of the side chains. All molecular mechanics calculations were carried out using the ICM software (12). All statistical analyses for differences were performed using Fisher’s exact test.

RESULTS AND DISCUSSION

Proteins with Collagen-related Structural Motifs in Bacteria and Viruses—The characteristic primary structure of collagen (glycine-X-Y repeats) and the high content of proline allowed us to construct a simple pattern (GPP)n for in silico detection of CSMs. The pattern was applied to 56 bacterial proteomes to detect such sequences. In addition, BLAST searches against 137 eubacterial genomes, 15 archeabacterial genomes, 30 ge-
A.

| Organisms          | no of CSMS | no of genomes | no of CSMS/ genome |
|--------------------|------------|---------------|-------------------|
| Viruses/phages     | 47 (47)    | *             | 1                 |
| Eubacteria         | 53 (25)    | 136           | 1-9               |
| Archaeabacteria    | 0          | 15            | 0                 |
| unicellular eukarya| 3 (3)      | 30            | 1                 |

B.

|                        | G          | X         | Y         |
|------------------------|------------|-----------|-----------|
| Human collagens        | 100%       | P 28.2%   | P* 38.5%  |
|                        |            | T 1.2%    | T 2.5%    |
|                        |            | Q 3.0%    | Q 6.0%    |
| Bacterial CSMSs        | 100%       | P 19.0%   | P 4.2%    |
|                        |            | T 0.4%    | T 31.6%   |
|                        |            | Q 1.7%    | Q 11.7%   |
| Viral/phage CSMSs      | 100%       | P 34.5%   | P 12.5%   |
|                        |            | T 1.0%    | T 18.1%   |
|                        |            | Q 0.7%    | Q 15.7%   |

Fig. 1. Occurrence and amino acid composition of CSMSs. A, the number of proteins containing a CSM in the various organisms is given, and the number of genomes by which these CSMSs are encoded is indicated within parentheses. The asterisk indicates that searches were made against all entries containing viral sequences. The number of genomes analyzed and the number of proteins containing CSMSs in a given genome are also shown. B, the position-specific distribution of glycines, prolines, glutamines, and threonines of the bacterial and viral CSMSs and human collagens is given. The values represent the percentage of a given amino acid in a specific position of the GXY triplet. The asterisk indicates that prolines in the Y position of human collagens are most often hydroxylated. Data on amino acid frequency for human collagens were from a previous report (32).

Fig. 2. Detailed view of the homology model of a threonine-rich CSM from B. anthracis. The conformation of threonine side chains was determined by an exhaustive systematic search procedure. Hydrogen bonds are shown in black. Part of the hydroxyproline-containing triple-helical structure in Protein Data Bank entry 1cag (33) is shown in cyan for comparison.

ruses, these were pooled for the purpose of subsequent analyses. Perhaps most striking is the difference in proline content and the preference for proline in position X in the CSMSs from bacteria and viruses. The proline content is significantly lower in these CSMSs as compared with human collagens \((p = 3.7 \times 10^{-147})\) for the bacterial CSMSs and \(p = 2.8 \times 10^{-18}\) for the viral CSMSs. In bacterial and viral CSMSs the fraction of prolines found in the Y position was significantly lower as compared with this fraction in human collagens \((p = 3.2 \times 10^{-77}\) for bacterial and \(p = 1.2 \times 10^{-39}\) for the viral CSMSs). Prolines in position Y are generally hydroxylated in human collagen; thus, the relative absence of prolines in this position among CSMSs from bacteria and viruses probably reflects their inability to synthetize hydroxyproline. In this context it should be mentioned that bacterial homologues of eukaryotic prolyl hydroxylases could not be identified despite extensive similarity searches.

The CSMSs identified contained a significantly higher proportion of threonine \((p = 7.6 \times 10^{-206} \text{ for bacterial CSMSs and } p = 1.9 \times 10^{-36} \text{ for viral CSMSs})\) and glutamine \((p = 2.1 \times 10^{-12} \text{ for bacteria and } p = 4.1 \times 10^{-14} \text{ for viruses})\) in the Y position as compared with human collagen. A minority of the bacterial CSMSs (17 of 53) had more than 50% of Y positions occupied by threonine (mean 94%). In these threonine-rich CSMSs, the X position is typically occupied by prolines, alanines, or serines, whereas charged residues are less common than in the other bacterial CSMSs and in human collagen. The proteins containing threonine-rich CSMSs are clustered in five bacterial species: the spore-forming human pathogens Clostridium difficile, Bacillus anthracis, and Bacillus cereus, the nitrogen-fixing Methylobacterium loti, and the sulfur-metabolizing Desulfitobacter hafniense. To determine whether threonine can influence stability of a collagen-like triple-helix, we built homology models of five representative proline-rich CSMSs on a trimeric \((\text{GPP})_{10}\) template constructed from a 1.3-A x-ray structure (11). The resulting models show that threonine in the Y position is able to form direct interchain hydrogen bonds to backbone carboxyls in its energetically most favored conformation (Fig. 2). When the amino acids in the X and Y positions of one threonine-rich CSMS were switched, the threonines were not able to form interchain hydrogen bonds (data not shown). Interestingly, there are indications that threonine in the Y position can also stabilize a triple-helical structure through indirect hydrogen bonding or through glycosylations (8, 13, 14).

Most of the remaining 83 CSMSs could instead be classified as...
A. Amino acid composition of bacterial and viral CSMs. A, the CSMs identified in bacterial and viral proteins (53 and 47, respectively) were categorized into three groups. The percentage of CSMs falling into each group is given. The threonine-rich CSMs have more than 50% of Y positions occupied by threonine and the proline-rich have more than 30% of X and Y positions occupied with proline, whereas charged CSMs have more than 45% charged residues in the X and Y positions. Some CSMs cannot be classified into one of these groups, whereas some meet more than one criterion. B, the skew in charge distribution between the X and Y position of GXY triplets from charged CSMs is shown. For every charged CSM, the net charge ($K/R = \pm 1$, $D/E = -1$) of residues in the Y position minus the net charge of residues in the X position was divided by the total number of charged residues. Thus, a CSM with all positively charged residues in the Y position and all negatively charged residues in the X position will obtain a value of +1, whereas a CSM with all positively charged residues in the X position and all negatively charged residues in the Y position will obtain a value of −1. Each dot in the diagram represents one CSM.

Bacterial and Viral Collagen-related Structural Motifs

---

| Charged (%) | Phage/virus (%) |
|-------------|-----------------|
| Threonine-rich | 32 | 19 |
| Proline-rich | 19 | 9 |
| None of above | 8 | 4 |
| >one of above | 4 | 6 |

---

**Evolutionary Aspects**—Because so few bacterial and unicellular eukaryotic genomes encode CSMs and because archaeabacteria completely lack such sequences, it seems unlikely that CSMs have evolved before the diversification of archaea, bacteri, and eukaryotes. It appears more plausible that genes encoding CSMs have moved horizontally or arisen on several occasions during evolution rather than selective gene loss among archaea, bacteria, and lower eukaryotes. Horizontal gene transfer between bacteria and vertebrates has been proposed to be relatively common (29, 30), although this view has also been questioned (31). If horizontal gene transfers of CSMs have occurred, we believe that the direction of transfer has been from eukaryotes to bacteria. After horizontal transfer of sequences encoding a CSM, bacteriophages may have promoted further horizontal transfer within the bacterial kingdom leading to the “patchy” distribution seen for CSMs among bacteria. We have no evidence for a role of viruses or bacteriophages in the putative horizontal transfer between eukaryotes and bacteria, but this cannot be excluded. After the putative horizontal transfer, extensive rearrangements seem to have occurred in the bacterial CSMs, because there are limited sequence similarities except for the spacing of glycines in the bacterial CSMs.

The limited sequence similarities also made phylogenetic analyses meaningless. Interestingly, the threonine-rich CSMs are very dissimilar in amino acid composition from human collagens and from the other bacterial CSMs. Instead, the threonine-rich CSMs have their closest homologues in hypothermal vent worm cuticle collagens (14). This suggests that several events of horizontal gene transfer of the genetic material encoding CSMs could have occurred during evolution. An alternative evolutionary explanation to the threonine-rich CSMs is that these have arisen de novo in bacteria, followed by horizontal spread to a few other bacterial species. The threonine-rich CSMs are highly repetitive and could have evolved from duplications of a short DNA segment.

The presence of CSMs in bacteria and viruses underlines the importance of collagen as a structural motif in nature. This work also suggests that alternative means for triple-helix stabilization probably operate in bacteria and viruses, and future elucidation of the structure and function of bacterial and viral CSMs represents an interesting scientific task.
Acknowledgments—The Institute of Genomic Research and other generous providers of sequence data are acknowledged. We thank Dr. Robert Janulczzyk for constructive comments.

REFERENCES
1. Beck, K., and Brodsky, B. (1998) J. Struct. Biol. 122, 17–29
2. Charalambeus, B. M., Keen, J. N., and McPherson, M. J. (1988) EMBO J. 7, 2903–2909
3. Bamford, D. H., and Bamford, J. K. (1990) Nature 344, 497
4. Medveczky, M. M., Geck, P., Vassallo, R., and Medveczky, P. G. (1993) Virus Genes 7, 349–365
5. Smith, M. C., Burns, N., Sayers, J. R., Sorrell, J. A., Casjens, S. R., and Hendrix, R. W. (1998) Science 279, 1834
6. Ruggiero, F., Exposito, J. Y., Bournat, P., Gruber, V., Perret, S., Comte, J., Olagnier, B., Garrone, R., and Theisen, M. (2000) FEBS Lett. 469, 132–136
7. Perret, S., Merle, C., Bernocco, S., Berland, P., Garrone, R., Hu, D. J. S., Theisen, M., and Ruggiero, F. (2001) J. Biol. Chem. 276, 43693–43698
8. Bann, J. G., Peyton, D. H., and Bachinger, H. P. (2000) FEBS Lett. 473, 237–240
9. Yang, W., Chan, V. C., Kirkpatrick, A., Ramshaw, J. A., and Brodsky, B. (1997) J. Biol. Chem. 272, 28837–28840
10. Xu, Y., Keene, D. R., Bujnicki, J. M., Hook, M., and Lukomski, S. (2002) J. Biol. Chem. 277, 23127–23134
11. Berisio, R., Vitagliano, L., Mazzarella, L., and Zagar, A. (2002) Protein Sci. 11, 262–270
12. Abagyan, R. A., and Totrov, M. M. (1994) J. Mol. Biol. 235, 983–1002
13. Kramer, R. Z., Bella, J., Mayville, P., Brodsky, B., and Berman, H. M. (1999) Nat. Struct. Biol. 6, 454–457
14. Mann, K., Mechling, D. E., Bachinger, H. P., Eckerskorn, C., Gaill, F., and Timpl, R. (1996) J. Mol. Biol. 261, 255–266
15. Chan, V. C., Ramshaw, J. A. M., Kirkpatrick, A., Beck, K., and Brodsky, B. (1997) J. Biol. Chem. 272, 31441–31446
16. Doi, T., Higashino, K., Kuribara, Y., Wada, Y., Miyazaki, T., Nakamura, H., Uesugi, S., Imanishi, T., Kawabe, Y., and Bakura, H. (1993) J. Biol. Chem. 268, 2126–2133
17. Sylvester, P., Couture-Tosi, E., and Mock, M. (2003) J. Bacteriol. 185, 1555–1563
18. Caldentey, J., Tuma, R., and Bamford, D. H. (2000) Biochemistry 39, 10566–10573
19. Erickson, P. R., and Herzberg, M. C. (1987) J. Immunol. 138, 3360–3366
20. Rasmussen, M., Edén, A., and Björck, L. (2000) Infect. Immun. 68, 6370–6377
21. Lukomska, S., Nakashima, K., Abdi, I., Cipriano, V. I., Ireland, R. M., Reid, S. D., Adams, G. G., and Musser, J. M. (2000) Infect. Immun. 68, 6542–6553
22. Whatmore, A. (2001) Microbiology 147, 419–429
23. Lukomska, S., Nakashima, K., Abdi, I., Cipriano, V. J., Shelvin, B. J., Graviss, E. A., and Musser, J. M. (2001) Infect. Immun. 69, 1729–1738
24. Rasmussen, M., and Björck, L. (2001) Mol. Microbiol. 40, 1427–1438
25. Sylvester, P., Couture-Tosi, E., and Mock, M. (2002) Mol. Microbiol. 45, 169–178
26. Erickson, P. R., and Herzberg, M. C. (1993) J. Biol. Chem. 268, 1646–1649
27. Duplessis, M., and Moineau, S. (2001) Mol. Microbiol. 41, 325–336
28. Janulczyk, R., and Rasmussen, M. (2001) Infect. Immun. 69, 4019–4026
29. Lander, E. S., Linton, L. M., Birren, B., Nussbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Leebachczyk, J., LeVine, R., McKean, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Songcz, M., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscou, R., et al. (2001) Nature 409, 860–921
30. Syvanen, M. (2002) Trends Genet. 18, 245–248
31. Stanhope, M. J., Lupas, A., Italia, M. J., Koresue, K. K., Volker, C., and Brown, J. R. (2001) Nature 411, 940–944
32. Ramshaw, J. A. M., Shah, N. K., and Brodsky, B. (1998) J. Struct. Biol. 122, 86–91
33. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Science 266, 75–81
Genome-based Identification and Analysis of Collagen-related Structural Motifs in Bacterial and Viral Proteins
Magnus Rasmussen, Micael Jacobsson and Lars Björck

J. Biol. Chem. 2003, 278:32313-32316.
doi: 10.1074/jbc.M304709200 originally published online June 3, 2003

Access the most updated version of this article at doi: 10.1074/jbc.M304709200

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 33 references, 13 of which can be accessed free at http://www.jbc.org/content/278/34/32313.full.html#ref-list-1