PROTEINS: Structure, Function, and Genetics 44:119–122 (2001)

Silk Fibroin: Structural Implications of a Remarkable Amino Acid Sequence

Cong-Zhao Zhou,[1,2,3] Fabrice Confalonieri,[1] Michel Jacquet,[1] Roland Perasso,[2] Zhen-Gang Li,[3] and Joel Janin[4]*
[1]*Institut de Génétique et Microbiologie, Université Paris-Sud et CNRS, Orsay Cedex, France*
[2]*Laboratoire de Biologie Cellulaire 4, Université Paris-Sud et CNRS, Orsay Cedex, France*
[3]*Department of Biology, University of Science and Technology of China, Hefei, Anhui, People's Republic of China*
[4]*Laboratoire d'Enzymologie et Biochimie Structurales, CNRS, Gif-sur-Yvette, France*

ABSTRACT **The amino acid sequence of the heavy chain of *Bombyx mori* silk fibroin was derived from the gene sequence. The 5,263-residue (391-kDa) polypeptide chain comprises 12 low-complexity "crystalline" domains made up of Gly–X repeats and covering 94% of the sequence; X is Ala in 65%, Ser in 23%, and Tyr in 9% of the repeats. The remainder includes a nonrepetitive 151-residue header sequence, 11 nearly identical copies of a 43-residue spacer sequence, and a 58-residue C-terminal sequence. The header sequence is homologous to the N-terminal sequence of other fibroins with a completely different crystalline region. In *Bombyx mori*, each crystalline domain is made up of subdomains of ~70 residues, which in most cases begin with repeats of the GAGAGS hexapeptide and terminate with the GAAS tetrapeptide. Within the subdomains, the Gly–X alternance is strict, which strongly supports the classic Pauling–Corey model, in which β-sheets pack on each other in alternating layers of Gly/Gly and X/X contacts. When fitting the actual sequence to that model, we propose that each subdomain forms a β-strand and each crystalline domain a two-layered β-sandwich, and we suggest that the β-sheets may be parallel, rather than antiparallel, as has been assumed up to now. Proteins 2001;44:119–122.**

Key words: *Bombyx mori*; low-complexity sequences; β-sheet packing; fiber structure

INTRODUCTION

The long and robust protein fiber that makes up the cocoon of the *Bombyx mori* caterpillar has been used by man for more than 3,000 years to produce high-quality thread and cloth. Their remarkable properties originate in the physical chemistry of silk fibroin, the protein that constitutes almost all the fiber. Fibroin is synthesized in large quantity by the silk gland of the caterpillar as two polypeptide chains linked by a disulfide bridge. The larger heavy chain is glycine rich, and most of its sequence is a repeat of Gly–Ala/Ser dipeptides. The silk fiber has been known as early as 1913 to diffract x-rays.[1] Its diffraction pattern is characteristic of a pleated β-sheet, and it helped Pauling and Corey in defining this type of secondary structure.[2] The Pauling–Corey model of silk fibroin[3] is revisited here within the context of the recently determined protein sequence of the fibroin heavy chain deduced from that of the genomic DNA.[4]

SEQUENCE AND COMPOSITION OF THE FIBROIN HEAVY CHAIN

After the early studies mentioned above and reviewed by Lucas et al.,[1] *Bombyx* silk fibroin received little attention from biochemists. The sequence of the 26-kDa light chain, of a short internal fragment and of a 85-residue C-terminal sequence of the heavy chain were determined.[5–7] The light chain, which is linked to the heavy chain by a single disulfide bridge, has a standard amino acid composition and a nonrepetitive sequence. It plays only a marginal role in the fiber.

The complete amino acid sequence of the heavy chain has now been deduced from that of a 80-kbp fragment of *Bombyx mori* genomic DNA comprising the 17-kpb fibroin gene (access code GenBank AF226688–EMBL P05790).[4] The gene codes for a polypeptide chain of 5,263 residues with a molecular weight of 391 kDa, composed of 45.9% glycine, 30.3% alanine, 12.1% serine, 5.3% tyrosine, 1.8% valine, and only 4.7% of the other 15 amino acid types. Most of the sequence is low-complexity and forms 2,377 repeats of a Gly–X (GX) dipeptide motif. The GX repeat is the building block of the β-sheets in the Pauling–Corey model and of the whole fiber. The sequence contains very long stretches of GX repeats that must constitute the x-ray diffracting structure, often called the "crystalline" component of silk fibroin. Residue X is Ala in 64% of the repeats, Ser in 22%, Tyr in 10%, Val in 3%, and Thr in 1.3% of the repeats. Essentially none of the other 14 amino acid types is present in the repeats. In 2% of the dipeptides, the first position is an alanine instead of glycine. This occurs almost exclusively in the GAAS tetrapeptide, which is repeated 41 times. Moreover, two hexapeptides s = GAGAGS, in 433 copies, and y = GAGAGY, in 120 copies, account for 70% of the low-complexity region. Their abundance was known from partial sequences and it was suggested that fibroin might be made up entirely of these

Grant sponsor: Programme des Recherches Avancées Franco-Chinois; Grant number: BT97-05.

*Correspondence to: Joèl Janin, LEBS-CNRS, 91198 Gif-sur-Yvette, France. E-mail janin@lebs.cnrs-gif.fr

Received 3 November 2000; Accepted 5 March 2001

Fig. 1. The silk fibroin heavy chain. The 5,263-residue polypeptide chain (EMBL P05790) is broken in domains and subdomains. The sequence number of the first residue number and the length *l* of each domain are given in parentheses. The 25-residue motif in boldface characters is repeated between the header and the linkers. A one-residue (or three-residue) insertion in subdomain GX9.4 is also in boldface characters. Lowercase letters s, y, a, and u represent frequently observed hexapeptides. Hexapeptide code and number of copies:

s GAGAGS 433 y GAGAGY 120 a GAGAGA 27 μ GAGYGA 39

hexapeptides.[6,8] Figure 1 shows that this is incorrect in spite of the hexapeptide abundance, and that fibroin has a more elaborate primary structure as discussed below.

THE NONREPETITIVE OR "AMORPHOUS" SEGMENTS

The GX repeats that form the bulk of the polypeptide chain are distributed among 12 domains separated by short linkers (Fig. 1). In contrast to the domains, the N-terminal 151 residues, C-terminal 50 residues, and the 42–43 residue linkers between domains are nonrepetitive and "amorphous," as opposed to the "crystalline" domains. The N-terminal segment or header has a standard amino acid composition and may constitute an independent globular unit, possibly with some α-helix as well as a β-sheet. Related sequences (33–38% identity) are found at the N-terminus of the fibroins of the moths *Galleria mallonella* and *Antheraea pernyi* (Fig. 2). The cocoon of the latter is used in Asia to produce tussah silk. *Antheraea* fibroin is otherwise completely different from *Bombyx*. It yields a different type of fiber, and the bulk of its 2,639-residue sequence (EMBL O76786), although low-complexity, is more alanine than glycine-rich.[9] A BLAST search against the *Drosophila* genome also suggests the presence of a segment related to these N-terminal sequences in the CG18026 gene product, but the score is much lower (27% identity over 136 residues), and the protein function is unknown.

The 58-residue C-terminal segment of the *Bombyx* heavy chain is arginine/lysine-rich and depleted in hydrophobic residues, which does not suggest a globular fold. It contains three cysteine residues involved in two disulfide bonds,[7] one with the light chain, the other internal. The 11 linker segments connecting the crystalline domains have nearly identical sequences, including a 25-residue nonrepetitive peptide (boldface characters in Fig. 1), also present in the header sequence in a truncated version. The peptide breaks the GX alternance and terminates the crystalline

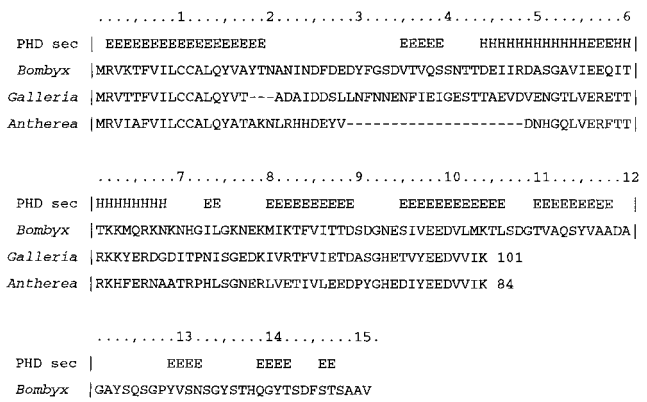

Fig. 2. The header sequence. Secondary structure prediction by PHD and an alignment of the N-terminal sequences of silk fibroins from *Bombyx mori* (EMBL P05790), *Galleria mellonella* (EMBL AF095239), and *Antherea pernyi* (EMBL O76786). E, extended; H, helical.

domains. It has a proline, charged residues and the only tryptophan residue found in fibroin. Charged residues are entirely absent from the crystalline domains.

PUNCTUATIONS IN THE CRYSTALLINE REGION

The 12 crystalline domains are labeled GX1–GX12 in Figure 1. Their average length is 413 residues, omitting GX12, which is much shorter (37 residues). Within a domain, the GX alternance is perturbed only by the occasional presence of a GAAS tetrapeptide, and, in domain GX9, by a single residue insertion at position 4108. The insertion changes the phase of the alternance and must perturb the β-sheet packing. In contrast, the GAAS tetrapeptide maintains the phase while introducing a punctuation in the crystalline domains. In the gene, all GAAS tetrapeptides use the same codons forming the same 12-base pair (bp) element.[4] Taking GX2 as an example, we find that the 511 residues of this domain are distributed into six subdomains beginning with a stretch of s = GAGAGS hexapeptides and ending in GAAS. The six subdomains are obviously related to each other, and it is most likely that the whole domain derives from successive duplications.

The same pattern of subdomains and duplications is seen in domains GX2 to GX11, except that GAGS or GTGS sometimes replace GAAS as a punctuation (Fig. 1). In the gene, these tetrapeptides all have a distinctive codon usage.[4] Domain GX1 can also be viewed as being made of subdomains ending in GAAS, but their sequence is divergent and four of them lack GAGAGS repeats. In total, the heavy chain comprises some 64 subdomains, each ~70 residues in length. Remarkably, the 70-residue segmentation of the protein sequence also shows up at the DNA level, where a repeating sequence unit of ~208 bp is observed, possibly the size of the DNA in a nucleosome.[4]

A MODEL OF THE SILK STRUCTURE: PAULING AND COREY REVISITED

Based on the diffraction pattern and the peculiar amino acid composition of silk, Pauling and Corey and their collaborators proposed that the fiber is made of antiparallel β-sheets packing on top of each other.[3] The β-strands extend along the fiber axis, yielding a 7.0-Å axial repeat containing two peptide units. There is also a ≈9.5-Å repeat across the fiber in the diffraction pattern. It is interpreted as representing in one direction, twice the spacing between β-strands in an antiparallel β-sheet, and in the orthogonal direction, twice the spacing between packed β-sheets. The Pauling–Corey model, elaborated upon by Crick and Kendrew,[10] takes note of the fact that, in a β-strand made of Gly–Ala/Ser repeats, all the nonglycine residues have their side-chains on the same face of the β-sheet. The β-sheets may then pack on each other alternatively by their glycine face and by their Ala/Ser face. The first packing yields a short 3.5–3.9 Å spacing, the second longer one of 5.3–5.7 Å. This type of packing was later observed in crystalline poly-(Ala–Gly).[11]

The Pauling–Corey model was challenged in a recent study of the *Bombyx* silk fiber pattern. In an attempt to interpret the diffracted intensities quantitatively, Takahashi et al.[12] test four ways of assembling β-strands and packing β-sheets. Adjacent strands within a β-sheet may either be parallel or antiparallel, and their side-chains may either all point in the same direction (polar mode) or alternatively up and down (antipolar mode). With this convention, the Pauling–Corey model is polar–antiparallel. It yields the best fit to a set of 26 diffracted intensities assuming the fibroin sequence to be just poly-(Ala–Gly). Nevertheless, the authors go on to consider effects of disorder and of the presence of serine residues on the calculated diffraction pattern of other types of assemblies. Their conclusions favor an antipolar–antiparallel model with Ala/Ser side-chains pointing alternatively up and down across a β-sheet. As antipolar β-sheets do not have a glycine-only face, their packing is less regular than in the Pauling–Corey model. The fit to the diffraction data depends on a number of assumptions, on the amino acid sequence for instance, and none of the alternative models can be entirely excluded.

FOLDING THE HEAVY CHAIN

Folding a 5,263-residue polypeptide chain must take place in several steps, whether conceptually or in reality. The actual sequence forbids fitting the packed β-sheet model to the whole fibroin molecule, but the model is plausible for the crystalline domains, which are large units able to form several β-sheets each. We suggest that each subdomain forms a β-strand, ~66 residues or 200 Å long, connected to the next β-strand by a four-residue β-turn at the GAAS boundary tetrapeptide. These β-strands are much longer than in globular proteins, where they rarely exceed 15 residues, but the strong diffraction pattern of silk suggests long β-strands. A crystalline domain will then comprise either one β-sheet or two β-sheets packing on top of each other. The latter is more likely in GX1, GX2, GX5, GX7, GX8, and GX11, which contain six or more subdomains. Each one of these domains could constitute a structural unit made of two layers of three-stranded

β-sheets, with approximate dimensions of 10 × 15 × 200 Å. A thicker and broader structure can then be created by packing domains side by side and layer by layer both within and between fibroin chains.

The next question is the polarity of the β-strands and β-sheets. The strict GX alternance observed within subdomains implies an extreme selective pressure against substituting glycines. This is more easily understood in the polar mode, where the β-sheets have one face devoid of side-chains, than in a nonpolar mode, where there are side-chains on both faces. On the other hand, the sequence has an obvious N-to-C directionality, with GAGAGS repeats at the N-terminus of the subdomains, and the rarer tyrosine, valine and threonine residues near the C-terminus. In an antipolar mode of β-sheet assembly, no regular packing of the large side-chains of these residues can be envisaged. In an antiparallel β-sheet, the large side-chains of one β-strand are at one end of the β-sheet, those of adjacent β-strands, at the opposite end. On the other hand, a polar–parallel assembly has all the GAGAGS hexapeptides at one end of the β-sheet, all the large side-chains clustering together at the other end, and a regular packing can be expected.

A natural way to fold a protein into parallel β-sheets is to build a two-layered solenoid structure, with successive β-strands alternating between the top and bottom layer. This type of assembly, which minimizes the conformation search made by the folding polypeptide chain, is common in globular all-β proteins. The transverse repeat of two β-strands observed in silk fibroin diffraction argues against parallel β-sheets, because their repeating unit is a single β-strand. However, only poly-(Ala–Gly) has a one-strand repeat, strictly speaking. In the actual sequence of the protein, the spatial distribution of the larger side-chains of serine and tyrosine may explain the larger unit cell. The fibroin diffraction pattern has weak reflections at low angles that cannot be indexed on the poly-(Ala–Gly) unit cell and some are attributed to the presence of serine in GAGAGS hexapeptides.[12]

CONCLUSIONS

The Gly–X alternance is a remarkable feature of the sequence of the *Bombyx mori* silk fibroin heavy chain, where it is maintained over the long stretches that constitute the crystalline domains. Whereas this alternance is the basis of the Pauling–Corey model, the model cannot easily be fitted as the actual sequence shows an additional level of repetition above the Gly–X unit, the subdomain, and a marked N-to-C directionality. Building a realistic model of the crystalline domains and of their association into a fiber, will require far more experimental information than can be derived from the fiber diffraction pattern alone.

REFERENCES

1. Lucas F, Shaw JBT, Smith SG. The silk fibroins. Adv Protein Chem 1958;13:108–244.
2. Pauling L, Corey RB. Proc R Soc Lond 1953;B141:21
3. Marsh RE, Corey RB, Pauling L. Biochim Biophys Acta 1955;16: 1–34
4. Zhou CZ, Confalonieri F, Medina N, Zivanovic Y, Esnault C, Yang T, Jacquet M, Janin J, Perasso R, Li ZG. Fine organization of *Bombyx mori* fibroin heavy chain. Nucleic Acids Res 2000;28:2413–2419.
5. Yamaguchi K, Kikuchi Y, Tagaki T, Kikuchi A, Oyama F, Shimura S, Mizuno S. Primary structure of the silk fibroin light chain determined by cDNA sequencing and peptide analysis. J Mol Biol 1989;210:127–139.
6. Gage LP, Manning RF. Internal structure of the silk fibroin gene of *Bombyx mori*. I The fibroin gene consists of a homogeneous alternating array of repetitious crystalline and amorphous sequences. J Biol Chem 1980;255:9444–9450.
7. Tanaka K, Kajiyama N, Ishikura K, Waga S, Kikuchi A, Ohtomo K, Takagi T, Mizuno S. Determination of the site of the disulfide linkage between heavy and light chains of silk fibroin produced by *Bombyx mori*. Biochim Biophys Acta 1999;1432:92–103.
8. Mita K, Ichimura S, Zama M, James TC. Specific codon usage pattern and its implication on the secondary structure of silk fibroin mRNA. Mol Biol 1988;203:917–925.
9. Sezutsu H, Tamura T, Yukuhiro K. Characterisation of the full length fibroin gene of a wild silkworm *Antheraea pernyi* (submitted).
10. Crick FHC, Kendrew J. Adv Protein Chem 1957;12:160.
11. Lotz B, Brack A, Spack G. β-Structure of periodic copolypeptides of L-alanine and glycine. J Mol Biol 1974;87:193–203.
12. Takahashi Y, Gehoh M, Yuzuhira K. Structure refinement and diffuse streak scattering of silk (*Bombyx mori*). Int J Biol Macromol 1999;24:127–138.