J Mol Model (2006) 12: 829–834
DOI 10.1007/s00894-006-0107-1

ORIGINAL PAPER

Tarin A. Mason · Patrick J. McIlroy · Daniel H. Shain

Structural model of an antistasin/notch-like fusion protein from the cocoon wall of the aquatic leech, *Theromyzon tessulatum*

Received: 29 November 2005 / Accepted: 11 January 2006 / Published online: 8 March 2006

Abstract The aquatic leech, *Theromyzon tessulatum*, secretes a proteinaceous cocoon with extraordinary physical properties (e.g., proteolytic, thermal resiliency). The deduced amino acid sequence of a major protein (Tcp—*Theromyzon* cocoon protein) from the *T. tessulatum* cocoon wall has been used to model the endogenous structure of the Tcp protein. The Tcp protein sequence comprises six internal repeats, each containing 12 ordered Cys residues. Amino acid alignments suggest that the region Cys1→6 is homologous to antistasin, a leech anticoagulant, and Cys7→12 is homologous to an epidermal growth factor-like domain found in notch-class proteins, which play critical roles in development, signaling, and adhesion throughout the Animalia. Modeling of individual domains (i.e., antistasin and notch) positions multiple hydrophobic and charged residues on the surface. When the antistasin and notch domains were fused, hydrophobic pockets appeared that may facilitate a polymerization mechanism. Collectively, the predicted features of our Tcp model are consistent with the physical properties of the leech cocoon wall.

Introduction

Many invertebrates produce biomaterials with unique physical properties (e.g., silk fibroin, and dragline), some of which have commercial or industrial applications [1, 2]. To date, most substances examined from segmented worms (i.e., phylum Annelida) are limited to the contents of leech salivary glands and have various medical applications (e.g., analgesics and anticoagulants [3, 4]). Largely overlooked, however, is that clitellate annelids (i.e., oligochaetes and leeches) secrete cocoons with extraordinary physical properties. For example, the glossiphoniid leech, *Theromyzon tessulatum*, secretes a proteinaceous cocoon that is resistant to high temperature (>250°C), denaturing reagents (e.g., guanidinium isothiocyanate, β-mercaptoethanol, and urea), organic chemicals, and proteases [5]. Ultrastructural analyses indicate that leech cocoon walls comprise multiple layers of fibrillar arrays stacked upon each other at various step-angles [6, 7].

Tcp (*Theromyzon* cocoon protein), a novel cysteine (Cys)-rich (~18%) protein, constitutes a major protein component of the *T. tessulatum* cocoon wall [5]. The Tcp protein (408 amino acids in length) contains six tandem repeats, each comprising 12 ordered Cys residues in an ~62 amino acid repeating unit. The first six Cys residues in each repeat share sequence similarity with antistasin, a leech anticoagulant that displays proteolytic resistance [3] and whose 3D structure has been solved [8]. We show in this study that the latter half of each Tcp repeat is similar in structure to notch, a single-spanning transmembrane protein whose 3D structure is also known [9], and that linking the antistasin- and notch-like domains generates a predicted monomer with properties consistent with those observed in the *T. tessulatum* cocoon wall.

Materials and methods

Homology searches

GenBank [10], SWISS-PROT [11], BLAST (Basic Local Alignment Search Tool; [12]), Protein Data Bank [13], and Meta Server [14] searches were conducted on the entire Tcp open reading frame, and representative repeat and domain sequences. SWISS-PROT accession numbers appear with all homologues used in alignments. Dotter software [15] and the Sanger Institutes Pfam domain retrieval program [16] were employed to retrieve sequence domain repeats and protein family sequence similarities, respectively. Clustal-X [17] and GeneDoc [18] were used for alignment purposes.

T. A. Mason · P. J. McIlroy · D. H. Shain (✉)
Biology Department, The State University of New Jersey,
315 Penn Street, Rutgers,
Camden, NJ 08102, USA
e-mail: dshain@camden.rutgers.edu
Tel.: +1-856-2256144
Fax: +1-856-2256312

Modeling strategy

Sections of the Tcp repeat corresponding to known structures were modeled using the Swiss-Model Automated Comparative Protein Modeling Server ([19]; http://swissmodel.expasy.org/). Where necessary, disulfide linkages were established by loading the modeled Tcp peptide into CS Chem3DPro (CambridgeSoft, Cambridge, MA, USA), connecting the appropriate Cys residues and subjecting the peptide to energy minimization using the molecular mechanics methods available in the program.

Individual sections of the Tcp repeat in the model were joined using CS Chem3DPro. Identical amino acids in the carboxyl end of the amino peptide and in the amino end of the carboxyl peptide were overlaid. The peptide bond between the previous amino acid and the overlaid amino acid in each part of the repeat was broken, the resultant short chains discarded, and the two halves of the repeat were joined by creating the appropriate peptide bond.

Six consecutive repeats were joined as described above to generate a model of the known Tcp sequence using the Swiss-Model. During the process, some editing (both manual and computer-aided) of the files was necessary to maintain compatibility between the programs. Rasmol v2.7.2.1.1 [20, 21] and the Swiss-PDB-Viewer (http://www.expasy.org/spdbv/; [19]) were used routinely to examine and analyze the models generated (e.g., identification of hydrophobic pockets and charged residues).

Results

The six ~62 amino acid repeats that comprise the Tcp protein are characterized by 12 ordered Cys residues [5], which appear to form two epidermal growth factor (EGF)-like folds (Cys1→6 and Cys7→12, respectively; Fig. 1). BLAST [12] and other searches (3D-PSSM; [22]) found several structures that could serve as templates for modeling the structure of Tcp. The first half of each repeat (i.e., Cys 1→6 and intervening sequences) shares sequence similarity (BLAST e-value—0.024, Meta Server Rscore—36.2, and 33–44% sequence similarity depending on Tcp repeat), especially with regard to Cys spacing, to the protease inhibitor antistasin (1SKZ; P15358 [8]). Several notch homologues were found that could serve as templates for modeling the second half of each repeat (Cys 7→12), e.g., 1PB5; P46531 [9]. Of these, notch (P46531 [9]; Meta Server Rscore—2.41, 30–36% sequence similarity) was chosen because it provided the least number of gaps to align the six Cys residues. Other possible template sequences with EGF-like folds included basement membrane protein bm-40 (P09486, 1NUB_A [23]; Meta Server Rscore—2.88) and antistasin (1SKZ, P15358 [8]; Meta Server Rscore—2.18). Note that the spacing between Cys 3 and 4 in the notch domain is variable but the Cys cross-linking pattern is identical among notch homologues [9]; consequently, the disulfide cross-links in the Tcp Cys 7→12 domain (as with Cys1→6) can be predicted with

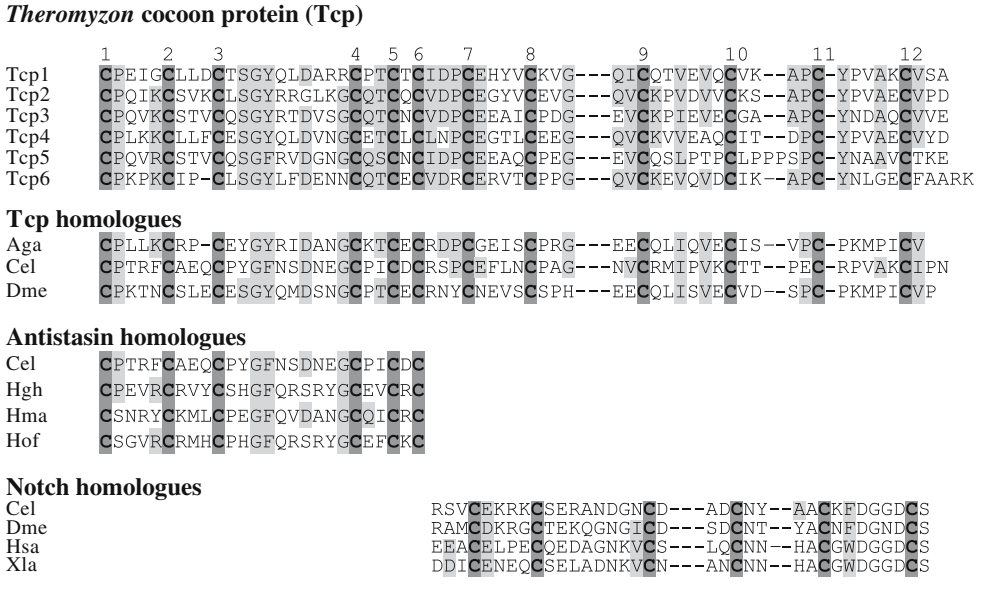

Fig. 1 Amino acid alignment of the *Theromyzon* cocoon protein (Tcp) Cys-rich repeats, antistasin, and notch sequences. Tcp 1–6 repeats (Q6QJ04) are shown in their entirety; ordered Cys residues are *numbered*, *aligned vertically* and *highlighted dark gray* while other conserved residues are *light gray* (based on Clustal-X [17]). *Dashes* indicate gaps introduced to maximize the sequence alignment. Tcp homologues are Aga (*Anopheles gambiae*—Q7QAA3), Cel (*C.elegans*—O44131), and Dme (*D. melanogaster*—Q9VB21). Antistasin homologues are Cel (*C.elegans*—O44131), Hgh (*Haementeria ghilianii*—P16242), Hma (*Hydra magnipapillata*—P38977), and Hof (*Haementeria officinalis*—P15358 [25]). Notch homologues are Cel (*C. elegans*—P14585), Dme (*D. melanogaster*—P07207), Has (*Homo sapiens*—P46531), and Xla (*Xenopus laevis*—P21783). The evolutionary relationships between Tcp repeats and homologues have been reported [5]

some certainty. Based on the respective templates, disulfide linkages were between Cys1–Cys4, Cys2–Cys5, and Cys3–Cys6 for the antistasin-like domain and Cys7–Cys11, Cys8–Cys10, and Cys9–Cys12 for the notch-like domain. It cannot be ruled out that free Cys residues are present in the endogenous Tcp protein, but this seems unlikely due to the conservation in Cys number and spacing in both antistasin- and notch-like domains.

The antistasin-like domain was constructed from the first half (Cys1→6) of the second repeat, plus three aa leader and trailer sequences (residues 93–124) using the published 3D structure of antistasin (residues 43–74, PDB file: 1SKZ [8]) and the alignment shown in Fig. 2. The notch-like domain was constructed from the second half (Cys7→12) of the second repeat, plus a three aa leader and one aa trailer (i.e., residues 122–155) using the published 3D structure of notch (residues 1447–1481, PDB file: 1PB5 [9]) and the alignment shown in Fig. 2. The predicted structures of the Tcp Cys1→6 and Tcp Cys7→12 domains are shown in Fig. 3. The second half of the fifth Tcp repeat, which contains two additional Pro residues between Cys 10 and 11 (see Fig. 1) and may represent a hypersensitive acid cleavage site [5], was overlaid with a representative Tcp repeat (i.e., Fig. 2b) and displayed modest deviation in its predicted backbone structure (Fig. 4a).

The two halves of the Tcp second repeat (i.e., Cys1→6 and Cys7→12) were joined as described in the "Materials and methods" section. Each of the amino acids in the overlapping tripeptide (Val, Asp, and Pro) was used independently for the overlay. A sample repeat structure is shown in Fig. 4. The three resulting structures, Val-, Asp- and Pro-overlaid repeats, were subsequently assembled into a putative Tcp-like structure by joining six repeats (i.e., Cys1→12) utilizing overlays of the first and last amino acids (Val) of the repeats. These structures were then examined for obvious flaws. Although differences were noted in the spatial relationship between the antistasin- and notch-like domains when the link was made at each of the three available positions, the Val-overlaid (Fig. 5) and Asp-overlaid (Fig. 6) structures generated extended arrangements consistent with the structure of naturally occurring EGF repeats [24], and also with consecutive antistasin repeats [8, 25]. The Pro-overlaid structure appeared sterically hindered and therefore was not considered further. Final models of the Val- and Asp-overlaid structures were constructed as described in the "Materials and methods" section, utilizing the multiple repeat structures described above.

```
                          1    2   3          4  5 6
Hof|0043-0074   RCTCSGVRCRMHCPHGFQRSRYGCEFCKCRLE
Tcp2|0093-0124  VSACPQIKCSVKCLSGYRFGLKGCQTCQCVDP
a

                         7    8        9      10   11     12
Hsa|1447-1481   EEACELPECQEDAGNKVCS---LQCNNHACGWDGGDCS
Tcp2|0122-0155 VDPCEGYVCEVG---QVCKPVDVVCKSAPC-YPVAECV
b
```

Fig. 2 **a** Tcp/antistasin and **b** Tcp/notch alignments used in homology modeling. Sequences corresponding to Tcp (repeat 2), antistasin (*H. officinalis* Cys1→6), and notch (*H. sapiens* Cys7→12) were aligned utilizing Clustal-X [17]. Cys residues are *numbered*, leader and trailer residues are *underlined*, *dark gray highlighting* indicates sequence identity and *light gray* indicates conservative amino acid differences

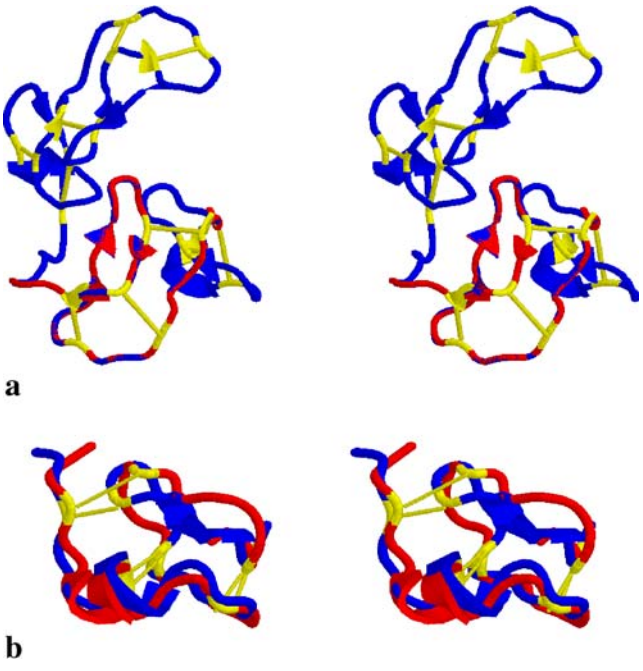

Fig. 3 Structural comparisons of Tcp with antistasin and notch domains. **a** Cross-eyed stereo view with overlaid structures of antistasin (1SKZ; *blue*), the Cys1→6 region of the second Tcp repeat (residues 93–124) as returned by Swiss-Model ([19]; *red*). **b** Cross-eyed stereo view with overlaid structures of notch (1PB5; *blue*), the Cys7→12 of the second Tcp repeat (residues 122–155) as returned by Swiss-Model ([19]; *red*). Disulfide bonds are *yellow*. These views were generated by RasMol [20, 21] with peptide backbones displayed using the "cartoons" option

The Val- and Asp-overlaid models presented structures that appeared to facilitate a polymerization mechanism. For example, the Val-overlaid repeat was an elongated "strand" characterized by a number of hydrophobic and charged residues on the surface (Fig. 7). In particular, valines in the outer boundary of the second fold (notch) protruded, and three of them formed a localized hydrophobic area. Rotating the structure by 180° about the *Z*-axis (Fig. 5) identified pockets and protrusions that could potentially interdigitate, facilitating hydrophobic and ionic interactions. This analysis generated a linear and somewhat zigzag-shaped monomer with a predicted a diameter of ~30 Å and length of ~120 Å.

The Asp-overlaid model generated a "lock washer" configuration (Fig. 6). The most notable characteristic was a large number of positively and negatively charged residues on the surface of the ring, providing a putative mechanism for stacks of rings to be held together by ionic and hydrophobic bonds. Hydrophobic residues also projected outwards in this model.

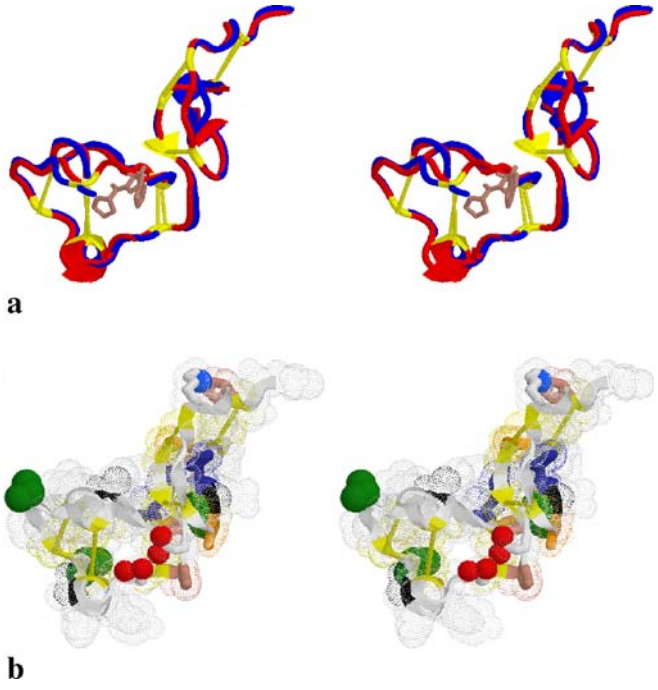

Fig. 4 Predicted molecular structure of a Tcp repeat (Cys1→12). Cross-eyed stereo views of the second Tcp repeat (residues 94–155) after joining the antistasin-like (Cys1→6) and notch-like (Cys7→12) domains utilizing a Val overlay. Disulfide bonds are indicated as bridges between the α-carbons (*yellow*). **a** Backbones only with the Tcp2 repeat displayed in *red* and the Tcp5 repeat in *blue*. A putative hypersensitive site comprising three consecutive Pro residues (331–333) is shown in flesh. **b** The predicted surface of molecule is displayed with the "dots" option. Identical and similar amino acid residues in the Tcp repeats (see Fig. 1) are as follows: Gly-108, Gly-133, and Ala-152 are *black* within the backbone; R-groups for Pro-97, Pro-124, and Pro-147 are indicated as *sticks* and *colored flesh*; R-groups for Tyr-109 and Tyr-149 are *sticks* and *colored mid-blue*; R-groups for Ser-107 and Thr-118 are *sticks* and *colored orange*; R-groups for Lys-100, Asp-123 and Glu-126 are *sticks* with the charged atoms as *spheres* and *colored blue* (+) or *red* (−); R-groups for Val-122, Val-135, and Val-139 are *green spheres*

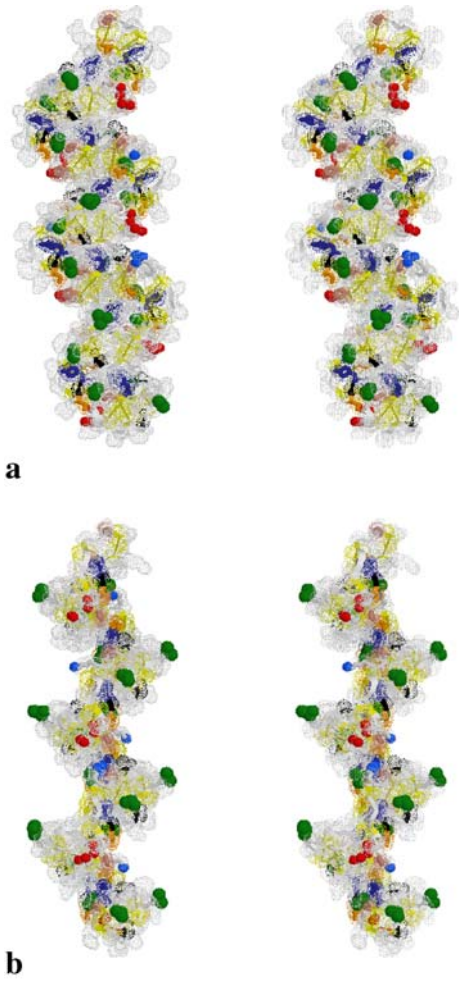

Fig. 5 Predicted structure of the Tcp "strand" monomer. Cross-eyed stereo view generated by linking six consecutive Tcp repeats (i.e., Cys1→12) using a Val overlay to join the antistasin and notch folds. These views were generated by RasMol [12, 13] with the peptide backbone displayed using the "cartoons" option, Cys–Cys bonds indicated as bonds between the α-carbons (*yellow*), and the surface of molecule displayed with the "dots" option. Conserved residues in the Tcp repeats are *colored* as in Fig. 4. The NH_2-terminus is at the *top* and (**b**) is a 90° rotation about the *vertical axis* of the view shown in (**a**)

Discussion

Tcp is a novel protein that assembles into a structure (i.e., leech cocoon) with extraordinary physical properties, hence our interest in predicting Tcp's native configuration. The results presented in this study suggest that endogenous Tcp acquires an extended arrangement with hydrophobic and charged residues projecting into solution. It is clear that outward protrusion is not thermodynamically stable for these residues (i.e., hydrophobic and charged); rather, they more likely associate with complementary residues, presumably from other monomers. In principle, Tcp monomers may polymerize to form the fibrous material that characterizes the cocoon wall of the leech [6, 7]. Individual fibers in the *T. tessulatum* cocoon are ~160 Å in diameter [7], and therefore Val-overlaid monomers would require lateral associations (possibly in a staggered arrangement) to form appropriately sized fibers; Asp-overlaid monomers (i.e., lock washer) were ~100 Å in diameter and therefore appear too small to build cocoon fibers. Other proteins may contribute to the fibrous composition of the leech cocoon (e.g., as in fibroin silk [2]), but there are currently no candidates for such a protein [5]. Regardless of the composition and the precise arrangement of Tcp monomers in the leech cocoon wall, however, the potential for hydrophobic associations between highly cross-linked Tcp monomers is consistent with the cocoon's ability to withstand extreme heat (>250°C [5]).

Tcp is not clearly associated with any classified protein family, and its high Cys content (17.8%) is a unique feature exceeded by only a few other reported proteins (e.g., hair [26]). In searching for domain sequence similarity with respect to other proteins, BLAST and ProSite identified several proteins with Cys repeat domains. Among the most well-characterized, Cys-rich domain proteins are those with EGF motifs, which are generally found in the

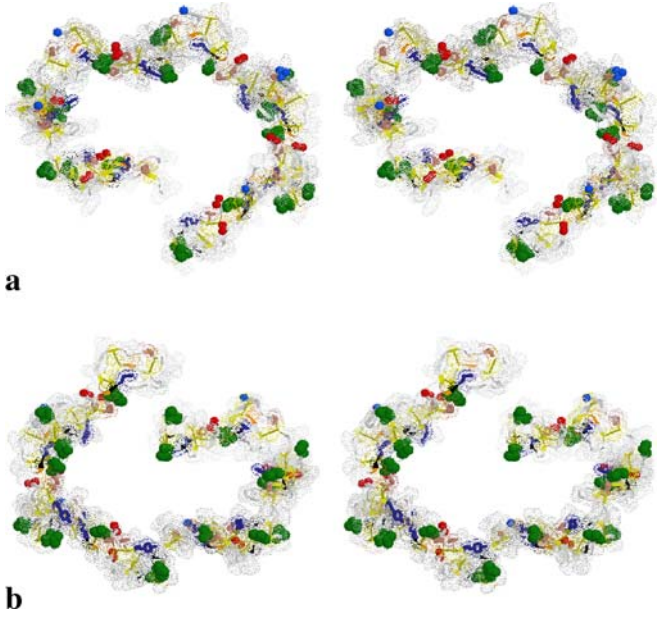

Fig. 6 Predicted structure of the Tcp "lock washer" monomer. Cross-eyed stereo view generated by linking six consecutive Tcp repeats (i.e., Cys1→12) using an Asp overlay to join the antistasin and notch folds. These views were generated by RasMol [12, 13] with *coloring* as in Fig. 4. (**b**) is a 90° rotation about the *horizontal axis* of the view shown in (**a**)

extracellular domains of membrane-bound or secreted proteins. The EGF domain includes six Cys residues that form disulfide bonds, and these are highly variable in their spacing and length. Subdomains between the conserved Cys in EGF-like proteins vary in length with at least one (and usually two) Gly residues between the fifth and sixth Cys residues [24], but this pattern was not observed in Tcp.

Evolutionary considerations

Tcp models were constructed by overlaying the first halves (Cys1→6) of Tcp1–Tcp6 using antistasin as the model backbone, and the second halves (Cys7→12) were modeled over notch backbones. It is interesting to note that the full Cys repeat of antistasin is annelid-specific and contains 10 ordered Cys residues, the first four of which align with Cys9→12 and the last six of which align with Cys1→6 in Tcp (see Fig. 1). The partial sequence similarity of antistasin to the putative notch-like domain in Tcp suggests a common and complex ancestry of these three Cys-rich proteins.

The notch proteins encompass a family of transmembrane receptors that have been highly conserved through evolution as mediators of cell fate (e.g., proliferation, cell survival, migration, and adhesion [27]). After cleavage of the full-length protein, notch is expressed as a heterodimeric receptor on the cell surface where it contacts its specific ligand (Delta) to begin the notch-signaling cascade. In comparison, endogenous notch has multiple domains including a stretch of 29–36 EGF motifs that, when deleted, causes notch to be degraded. Although no

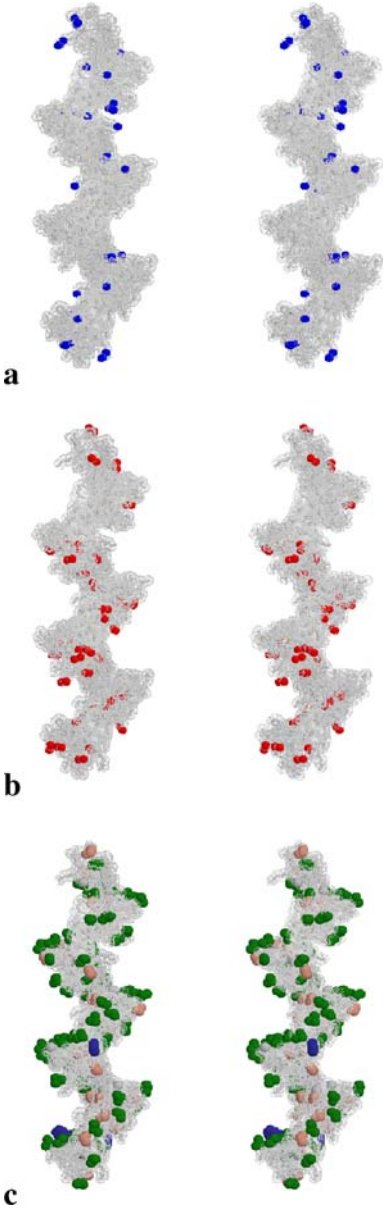

Fig. 7 Cross-eyed stereo views of "strand" highlighting features expected to be involved in polymerization. **a** Positively charged atoms (*blue*). **b** Negatively charged atoms (*red*). **c** Val, Leu, and Ile resides (*green*); Pro residues (*flesh*); Phe residues (*blue*)

specific EGF repeats appear to protect notch against proteolytic cleavage, this phenomenon appears to be dependent on a general feature of EGF motif structure [28]. It is notable that the *T. tessulatum* cocoon is resistant to proteolytic enzymes [5], and thus the repeating EGF-like domains in the Tcp protein may contribute to the proteolytic resiliency of the leech cocoon. Furthermore, only full-length notch 1 can form homodimers while EGF negative constructs appear to exist only as monomers, suggesting that EGF motifs facilitate dimerization [28], which consistent with our notion that Val- and/or Asp-overlaid structures have the capacity to polymerize.

Tcp has homologues in nematodes and arthropods (*Caenorhabditis elegans*, *Drosophila melanogaster*, and

Anopheles gambiae), all of which contain a conserved 12 Cys repeat, suggesting that a Tcp-like protein functioned in an ancestral metazoan. It is possible that antistasin- and notch-like domains were precursors of Tcp, and these were fused together by exon shuffling [29] and then duplicated in tandem to generate the full-length Tcp gene. Intervening residues between the ordered Cys have deviated from each other (i.e., compare structures of different Tcp repeats, Fig. 2), and these changes may specify the properties of Tcp-like proteins from different animal phyla. Nonetheless, some structural and functional aspects of individual domains (e.g., Cys cross-links, proteolytic resistance, and dimerization) appear to be conserved in endogenous Tcp and related proteins. The evolution of Tcp in leech (and probably in the annelid lineage) is perhaps best described as the fusion of two independent peptide domains (i.e., antistasin- and notch-like) that created a protein with novel, and potentially applicable, properties.

Acknowledgements We thank Sonia Krutzke for the editorial comments. This work was supported by NSF grant IBN-0417081000 to DHS.

References

1. Denny MW (1980) In: The Mechanical Properties of Biological Materials. Symposia of Society for Experimental Biology 34:247–272. Cambridge Univ. Press, Cambridge
2. Rui HG (1998) Silk reeling: cocoon silk study. Science
3. Tuszynski GP, Gasic TB, Gasic GJ (1987) J Biol Chem 262:9718–9723
4. Greinacher A (2004) Expert Rev Cardiovasc Ther 2:339–357
5. Mason TA, McIlroy PJ, Shain DH (2004) FEBS Lett 561: 167–172
6. Knight DP, Hunt S (1974) Comp Biochem Physiol 47A:871–880
7. Dimitriu C, Shain DH (2004) Micron 5:281–285
8. Schreuder H, Arkema A, de Boer B, Kalk K, Dijkema R, Mulders J, Theunissen H, Hol W (1993) J Mol Biol 231: 1137–1138
9. Vardar D, North CL, Sanchez-Irizarry C, Aster JC, Blacklow SC (2003) Biochemistry 42:7061–7067
10. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) Nucl Acids Res 33:34–38
11. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) Nucl Acids Res 31:3784–3788
12. Altschul SF, Gish W, Miler W, Myers EW, Lipman DJ (1990) J Mol Biol 3:403–410
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucl Acids Res 28:235–242
14. Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) Bioinformatics 22:1015–1018
15. Sonnhammer ELL, Durbin R (1995) Gene 167:1–10
16. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jnes S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR (2004) Nucl Acids Res 32:138–141
17. Higgins DG, Thompson JD, Gibson TJ (1996) Methods Enzymol 266:383–402
18. Nicholas Jr HB, Deerfield 2nd DW, Ropelewski AJ (2000) Biotechniques 28:1174–1178
19. Guex N, Peitsch MC (1997) Electrophoresis 18:2714–2723
20. Sayle R, Milner-White EJ (1995) Trends Biochem Sci 20:374–376
21. Bernstein HJ (2000) Trends Biochem Sci 25:453–455
22. Kelley LA, MacCallum RM, Sternberg MJE (2000) J Mol Biol 299:499–520
23. Sasaki T, Hohenester E, Gohring W, Timpl R (1998) EMBO J 17:1625–1634
24. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) Nucl Acids Res 30:276–280
25. Han JH, Law SW, Keller PM, Kniskern PJ, Silberklang M, Tung JS, Gasic TB, Gasic GJ, Friedman PA, Ellis RW (1989) Gene 75:47–57
26. Dekio S, Jidoi J (1989) J Dermatol 16:453–457
27. Karsan A (2005) Can J Physiol Pharmacol 83:14–23
28. Sakamoto K, Chao WS, Katsube K, Yamaguchi A (2005) Exp Cell Res 302:281–291
29. Patthy L (1991) Bioessays 13:187–192