Structural basis of sequence-specific collagen recognition by SPARC

Erhard Hohenester*†, Takako Sasaki‡§, Camilla Giudici‡, Richard W. Farndale¶ and Hans Peter Bächinger§||

*Department of Life Sciences, Imperial College London, London SW7 2AZ, UK.

‡Department of Molecular Medicine, Max-Planck-Institut für Biochemie, 82152 Martinsried, Germany.

§Department of Biochemistry and Molecular Biology, Oregon Health & Science University, Portland, OR 97239, USA.

¶Department of Biochemistry, University of Cambridge, Cambridge CB2 1QW, UK.

||Shriner's Hospital for Children, Portland Research Center, Portland, OR 97239, USA.

†To whom correspondence shoud be addressed. Dr Erhard Hohenester, Biophysics Section, Blackett Laboratory, Imperial College London, London SW7 2AZ, UK. Phone +44 (0)2075947701, Fax +44 (0)2075890191, E-mail: e.hohenester@imperial.ac.uk.

Data deposition: The atomic coordinates have been deposited in the PDB (entry 2v53).

Abstract

Protein interactions with the collagen triple helix play a critical role in collagen fibril formation, cell adhesion and signalling. However, structural insight into sequence-specific collagen recognition is limited to an integrin-peptide complex. A GVMGFO motif in fibrillar collagens (O denotes 4-hydroxyproline) binds three unrelated proteins: von Willebrand factor (VWF), discoidin domain receptor 2 (DDR2) and the extracellular matrix protein SPARC/osteonectin/BM-40. We report the crystal structure at 3.2 Å resolution of human SPARC bound to a triple-helical 33-residue peptide harbouring the promiscuous GVMGFO motif. SPARC recognises the GVMGFO motifs of the middle and trailing collagen chains, burying a total of 720 $Å^2$ of solvent-accessible collagen surface. SPARC binding does not distort the canonical triple helix of the collagen peptide. In contrast, a critical loop in SPARC is substantially remodelled upon collagen binding, creating a deep pocket that accommodates the phenylalanine residue of the trailing collagen chain ('Phe pocket'). This highly restrictive specificity pocket is shared with the collagen-binding integrin I-domains, but differs strikingly from the shallow collagen-binding grooves of the platelet receptor glycoprotein VI and microbial adhesins. We speculate that binding of the GVMGFO motif to VWF and DDR2 also results in structural changes and the formation of a 'Phe pocket'.

Collagen, the most abundant protein in vertebrates, is characterised by a triple-helical structure of three polypeptide chains containing repetitive glycine-X-X' triplets; the X and X' positions, respectively, are often occupied by the imino acids proline and 4-hydroxyproline (O). The 28 human collagens have numerous essential functions in tissue formation, stability and homeostasis, and mutations in collagen genes cause many human diseases (1). Collagens I-III, V and XI form supramolecular fibrils that lend mechanical stability to the vertebrate body. Normal collagen fibrillogenesis *in vivo* requires a number of globular proteins interacting with the triple helix, such as small leucine-rich repeat proteoglycans (2). Cellular interactions with collagen are mediated by a diverse group of transmembrane collagen receptors, including integrins, discoidin domain receptors (DDRs) and members of the immunoglobulin superfamily (3-5). While the structure of the collagen triple helix has been known for more than 50 years (6), protein-collagen interactions remain poorly understood at the atomic level.

Synthetic triple-helical peptides have been invaluable in mapping specific binding sites within collagen (7). The major integrin-binding site in fibrillar collagens I and II is a GFOGER motif (8, 9). The atomic details of this important interaction were revealed by a crystal structure of the integrin α2 I-domain bound to a 21-residue triple helical peptide, providing the only example to date of a vertebrate protein-collagen complex (10). More recent studies have identified a GVMGFO motif, conserved in collagens I-III and situated ~100 residues (~30 nm) upstream of the GFOGER motif, as a binding hotspot for three structurally and functionally distinct proteins: the plasma protein von Willebrand factor (VWF), whose interaction with collagen contributes to haemostasis (11); the receptor tyrosine kinase DDR2 (12); and the extracellular matrix protein SPARC (13).

SPARC (also called osteonectin or BM-40) is a small evolutionarily conserved glycoprotein that modulates cell-matrix interactions and collagen assembly (14, 15). SPARC is essential for embryo development in invertebrates (16, 17) and has been suggested to act as a chaperone for basement membrane collagen IV (17). Mice lacking SPARC develop early-onset cataract, lax skin and bone loss; these phenotype are likely to result, at least in part, from perturbed collagen fibrillogenesis or cell adhesion (14, 15). Human SPARC consists of an acidic 52-residue segment followed by a follistatin-like (FS) domain and an α-helical domain (EC) containing two unusual calcium-binding EF-hands and the collagen-binding site (18-20). Tissue-derived SPARC is processed by proteolytic cleavage in the EC domain, resulting in a ~10-fold increase in collagen affinity. This effect can be mimicked by recombinant deletion of helix C (ΔαC mutant) (20). To obtain insight into the molecular mechanism of collagen recognition by SPARC, we have determined the crystal structure of a complex between SPARC FS-EC ΔαC and a triple-helical collagen peptide containing the GVMGFO motif.

Results and Discussion

Structure of the SPARC-Collagen Complex. SPARC FS-EC ΔαC binds with high affinity to residues 564-590 of the human collagen α1(III) chain (corresponding to 397-423 of the triple-helical domain) (13). A 33-residue peptide containing this collagen III region preceded by three GPO repeats forms a stable triple-helix and a 1:1 complex with SPARC FS-EC ΔαC (data not shown). We determined the crystal structure of this

SPARC-collagen complex at 3.2 Å resolution by a combination of molecular replacement and heavy atom phasing (Table S1, Fig. S1).

The collagen peptide is a straight triple helix of ~95 Å length (Fig. 1). The three collagen chains are coiled around each other with the characteristic one-residue stagger, leading to the designation of a leading, middle and trailing chain (10). The triple-helical twist does not vary much along the collagen peptide: the GPO-rich N-terminus is close to an ideal 7_5 helical symmetry, whereas the remainder of the helix assumes a twist halfway between 7_5 and 10_7 symmetry (21, 22). SPARC is bound to the C-terminal half of the collagen peptide, burying 720 $Å^2$ of solvent-accessible collagen surface. Residues 21-23 of the middle chain and residues 17-27 of the trailing chain make contacts closer than 4.0 Å with SPARC; the leading chain does not contribute to SPARC binding (Fig. 2, Table S2). Thus, the SPARC-binding site observed in the complex structure spans four collagen triplets, GQOGVMGFOGPK (570-581 of the collagen α1(III) chain). In principle, SPARC could equally well bind to the leading and middle collagen chains, whose relative disposition is identical to that of the middle and trailing chains. As in the case of the integrin-collagen complex (10), the observed mode of binding is likely to be determined by crystal packing forces.

Only the SPARC EC domain interacts with the collagen peptide, although the FS domain comes close (4.5 Å) to V20 of the middle chain (Fig. 2*B*). The collagen-binding site of the SPARC EC domain is composed of the long αA helix and the adjacent αE- αF loop (Fig. 1). These two elements form a deep pocket that accommodates F23 of the collagen trailing chain (Fig. 2*C*, Table S2). This 'Phe pocket', the most notable feature of the SPARC-collagen interface, is delineated by several apolar SPARC residues (F146,

M150, W153, L242) and, suprisingly, a salt bridge between R149 and E246. The refined side chain conformation of the collagen F23 in the 'Phe pocket' is unusual and allows the phenyl ring of F23 to approach SPARC residue W153 edge-on, suggestive of a C-H···π hydrogen bond (23). A similar interaction may also be formed with the carboxylate group of E246.

SPARC residues W153 and L242 are also involved in the second major apolar SPARC-collagen contact, with M21 of the collagen middle chain (Fig. 2*D*, Table S2). Finally, F23 of the middle chain is situated atop SPARC residue M245 at the apex of the αE- αF loop. In addition to these apolar interactions, a total of six direct hydrogen bonds are formed between SPARC and the collagen peptide, five with the collagen main chain and one with the hydroxyl group of O24 of the middle chain. Notably, all polar residues at the rim of the 'Phe pocket' (R149, E246, the indole nitrogen of W153) are engaged in hydrogen bonding with the collagen peptide (Figs. 2*C* and 2*D*).

Previous studies have shown that the glycan attached to N99 in the FS domain modulates the affinity of SPARC binding to fibrillar collagens I and V (24, 25). In our structure, the *N*-acetylglucosamine moities attached to N99 are at >10 Å distance from the collagen peptide (Fig. 1), but the glycan may cause steric hindrance when SPARC binds to fibrillar collagens. Recently, an atomic model of the collagen I fibril structure has been reported, based on diffraction patterns of fibrils from rat tail tendon (26). Inspection of this structure shows that the SPARC-binding site is largely exposed on the flat surface of the fibril's gap zone (not shown). Even so, a certain amount of local unfolding (disorder) of the fibril surface would appear to be required for SPARC binding to occur.

Structural Changes within SPARC. Collagen binding is accompanied by substantial structural rearrangements within SPARC. Most notably, the crucial 'Phe pocket' is not present in free SPARC, but is only formed upon collagen binding by extensive remodelling of the αE- αF loop in the EC domain (Fig. 3). In the three different structures of the uncomplexed SPARC EC domain (18-20), this loop acts like a flap, with I243 and P244 plugging the pocket and burying most of the critical W153 side chain (Fig. 3*A*). In the complex with collagen, the αE helix is extended by two irregular turns (there are prolines at positions 237, 241 and 244), thus opening up the pocket and completing its rim by forming the R149-E246 salt bridge (Figs. 3*B* and 3*C*).

The SPARC-collagen crystals grown at pH 5.5 are missing a calcium ion in the second EF-hand (Fig. S2), but we are confident that the remodelling of the αE-αF loop is a consequence of collagen binding and not of incomplete calcium loading. Crucially, loss of one of the two calcium ions does not compromise the native structure of the EF-hand pair. The partially and fully calcium-loaded EF-hand structures can be superimposed with a root-mean-square deviation of 0.59 Å for 44 Cα atoms, whereas complete removal of calcium from SPARC is known to result in large conformational changes and abolish collagen binding (27). The integrity of the EF-hand pair in the partially loaded state appears to be maintained by hydrogen bonding between the calcium ligands and the C256-C272 disulphide bridge at the empty EF-hand. Besides the αE-αF loop, there are other differences between free and collagen-bound SPARC, including a reorientation of the FS domain and changes in the αA-αB loop (Fig. S2); however, these changes are

distant from the SPARC-collagen interface and likely to be the result of crystal packing forces.

Collagen binding to the integrin I-domain results in a conformational change that is linked to integrin signal transduction (10). Whether the structural changes within SPARC are functionally important is unknown. Computational docking of collagen triple helices is often used in the absence of experimental protein-collagen complex structures (11, 28, 29). Our observation that collagen binding causes substantial structural changes within SPARC cautions against the uncritical use of such docking procedures.

Agreement with Previous Mutagenesis. We previously showed that SPARC residues R149, N156, L242, M245 and E246 are important for binding to fibrillar collagen I and basement membrane collagen IV (20). All of these residues are now seen to make multiple interactions with the collagen peptide, with the exception of N156 (Figs. 2 and 3). This agreement is remarkable given the unexpected structural changes within SPARC upon collagen binding. The critical role of N156 demonstrated by mutagenesis could be explained if this residue were involved in a water-mediated hydrogen bond with collagen (water molecules are not resolved in our structure). Indeed, a canonical structural water molecule (30) is predicted to be present in an appropriate position on the trailing chain, bound between the peptide carbonyl group of G22 and the hydroxyl group of O24. Collagen triple helices are highly hydrated (30) and other structural water molecules may also contribute to the SPARC-collagen interaction.

Collagen Binding by Vertebrate and Invertebrate SPARCs. The collagen-binding residues of human SPARC are highly conserved in all vertebrate and invertebrate SPARC sequences (Fig. S3 and data not shown), suggesting that all SPARCs bind collagen in a similar manner as the human orthologue. Indeed, collagen binding by invertebrate SPARCs has been demonstrated by genetic (17) and biochemical (31) experiments. Lacking an equivalent of helix αB, invertebrate SPARCs have a much shorter connection between helices αA and αD (Fig. S3) and may therefore not be subject to the same proteolytic activation mechanism as mammalian SPARCs (20).

Prediction of SPARC-Binding Sites in Collagens I-IV. The detailed collagen sequence requirements for SPARC binding have not been probed biochemically (13). From our structure, we predict an absolute requirement for phenylalanine at position 23 of the trailing chain. The apolar nature of the 'Phe pocket' would appear to preclude binding of polar side chains, whereas aliphatic side chains would not be able to fully occupy the available space. We further predict a preference for apolar side chains at positions 21 and 23 of the middle chain, as well as at position 20 of the trailing chain. Finally position 24 of both chains should be hydroxyproline to allow (water-mediated) hydrogen bonding interactions with N156 and E246 (Figs. 2*C* and 2*D*).

We have used these structure-derived rules to predict further SPARC-binding sites in the fibrillar collagens I-III and in basement membrane collagen IV. Collagen I is a [α1(I)]$_2$α2(I) heterotrimer, and we assumed that the two α1(I) chains are adjacent (i.e. not leading and trailing). Collagens II and III are homotrimers. The major SPARC-binding site in the fibrillar collagens I-III, GVMGFO, is ~600 residues from the C-terminus of the

triple helix (13). The only other plausible site from sequence analysis is a GATGFO sequence (GAAGFO in collagen III) halfway between the GVMGFO motif and the C-terminus. Rotary shadowing electron micrographs indeed show a secondary SPARC-binding site in this region (13). The six distinct mammalian collagen IV chains assemble into three heterotrimers, of which the $[\alpha1(IV)]_2\alpha2(IV)$ heterotrimer is the most abundant; invertebrates have only one type of collagen IV heterotrimer (1, 32). There is experimental evidence that the α2(IV) chain in the human $[\alpha1(IV)]_2\alpha2(IV)$ heterotrimer is trailing (33). We have identified four putative SPARC-binding sites in the human $[\alpha1(IV)]_2\alpha2(IV)$ heterotrimer, two of which are almost perfect matches for the structure-derived consensus (Fig. 4). Rotary shadow electron micrographs indeed show multiple SPARC-binding sites along the human collagen IV triple-helix (34), but the data do not permit the precise locations to be determined. It would be interesting to approach this question using peptide libraries (7), but the synthesis of the required heterotrimeric collagen peptides is technically challenging.

Comparison with Other Collagen-Binding Proteins. The GVMGFO motif recognised by SPARC bears little similarity to the GFOGER motif recognised by integrins, yet the two proteins appear to achieve sequence specificity by similar means. SPARC uses a restrictive pocket to bind the phenylalanine side chain of the GVMGFO motif (Figs. *2B* and 2*C*). Similarly, the crucial glutamic acid side chain of the integrin-binding GFOGER is bound in a deep pocket, co-ordinating the divalent metal ion at the metal ion-dependent adhesion site (10). SPARC and the integrin I-domain bury a total of 720 $Å^2$ and 660 $Å^2$ of solvent-accessible collagen surface, respectively, with both proteins undergoing structural

changes upon collagen binding. One notable difference is that the collagen peptide is kinked in the complex with the integrin I-domain, but remains completely straight in the complex with SPARC.

VWF and DDR2 recognise the same GVMGFO motif as SPARC, and the phenylalanine side chain in collagen is critical for all three interactions (11, 12). It is thus reasonable to assume that SPARC, VWF and DDR2 share a common mode of collagen recognition, in particular with regard to the critical phenylalanine residue. Whether the three proteins compete for collagen in a physiological situation remains to be studied.

The collagen-binding A3 domain of VWF has been crystallised and its collagen binding site mapped by NMR, antibody blocking and mutagenesis experiments (35-38). The collagen-binding site delineated by these studies is a shallow groove, with no obvious 'Phe pocket'. We hypothesise that a 'Phe pocket' is formed upon collagen binding to the VWF A3 domain and that W982, which lies buried beneath the collagen-binding site, is part of this pocket. In support of this hypothesis, there is evidence for conformational variability in this region: the four crystallographically independent structures of the A3 domain show substantial differences in the vicinity of W982 (35, 36), and fluorescence experiments have shown that W982 exists in multiple rotamers (39).

The collagen-binding site of in the discoidin domain of DDR2 also is a shallow groove, into which a generic $(GPO)_n$ triple helix could be docked (29). How DDR2 achieves sequence specificity was not revealed by this study, however. A likely candidate for binding the phenylalanine side chain of the GVMGFO motif is the surface-exposed side chain of W52, whose mutation to alanine abolishes collagen binding (29). Given that

the collagen-binding groove of DDR2 is made up of several long loops, pronounced structural changes upon collagen binding are to be expected.

Platelet glycoprotein VI (GPVI) binds to certain GPO-rich collagen sequences (40). Structure determination of the GPVI ectodomain has revealed a shallow groove appropriate for binding a $(GPO)_n$ triple helix (28), but the structural basis for sequence-specific collagen binding remains unknown. Finally, a number of microbial collagen receptors bind triple-helical collagen without any apparent sequence specificity. The structure of *Staphylococcus aureus* CNA in complex with a $(GPO)_n$ collagen peptide has revealed a mode of binding termed the 'collagen hug' (41). The collagen triple helix is clamped by the two CNA domains, one of which accommodates the GPO repeats of the peptide in a relatively featureless and mostly hydrophobic groove.

In summary, our structure has revealed how the hydrophobic GVMGFO motif in fibrillar collagens is recognised by SPARC. An unexpected conformational change in SPARC creates a deep specificity pocket that binds the phenylalanine side chain of the GVMGFO motif. The same motif is also recognised by VWF and DDR2, and we predict that collagen binding to these proteins results in the formation of similar 'Phe pockets'.

Methods

Peptide Synthesis, Protein Expression and Complex Formation. The collagen peptide was prepared by solid-phase synthesis as described in *SI Materials and Methods.* SPARC FS-EC ΔαC was produced in human embryonic kidney 293 cells as described (20). The SPARC-peptide complex was formed by mixing protein and peptide in a ~1:1.5 ratio and purified by size exclusion chromatography as described in *SI Materials and Methods.*

Crystallisation, Data Collection and Structure Determination. The SPARC-collagen complex was crystallised as described in *SI Materials and Methods.* X-ray diffraction data were collected and processed as described in *SI Materials and Methods.* The structure of the SPARC complex was solved by a combination of molecular replacement and SIRAS techniques, and refined to an R-factor of 0.254 (R_{free} = 0.320) at 3.2 Å resolution, as described in *SI Materials and Methods.* Data processing, phasing and refinement statistics are listed in Table S1.

Acknowledgments. We thank Noemi Fukuhara, Federico Carafoli and the staff at beamline 10.1 at the SRS Daresbury for help with data collection; Peter Brick for help with SHARP; and Jordi Bella for help with collagen refinement and analysis. This work was supported by grants from the Wellcome Trust (E.H.) and from Shriner's Hospital for Children (H.P.B.). E.H. is a Wellcome Senior Research Fellow.

References

1. Myllyharju J, Kivirikko KI (2004) Collagens, modifying enzymes and their mutations in humans, flies and worms. *Trends Genet* 20:33-43.
2. Ameye L, Young MF (2002) Mice deficient in small leucine-rich proteoglycans: novel in vivo models for osteoporosis, osteoarthritis, Ehlers-Danlos syndrome, muscular dystrophy, and corneal diseases. *Glycobiology* 12:107R-116R.
3. Leitinger B, Hohenester E (2007) Mammalian collagen receptors. *Matrix Biol* 26:146-155.
4. Heino J (2007) The collagen family members as cell adhesion proteins. *Bioessays* 29:1001-1010.
5. Sweeney SM*, et al.* (2008) Candidate cell and matrix interaction domains on the collagen fibril, the predominant protein of vertebrates. *J Biol Chem* 283:21187-21197.
6. Brodsky B, Persikov AV (2005) Molecular structure of the collagen triple helix. *Adv Protein Chem* 70:301-339.
7. Farndale RW*, et al.* (2008) Cell-collagen interactions: the use of peptide Toolkits to investigate collagen-receptor interactions. *Biochem Soc Trans* 36:241-250.
8. Knight CG*, et al.* (2000) The collagen-binding A-domains of integrins α1β1 and α2β1 recognize the same specific amino acid sequence, GFOGER, in native (triple-helical) collagens. *J Biol Chem* 275:35-40.
9. Xu Y*, et al.* (2000) Multiple binding sites in collagen type I for the integrins α1β1 and α2β1. *J Biol Chem* 275:38981-38989.
10. Emsley J*, et al.* (2000) Structural basis of collagen recognition by integrin α2β1. *Cell* 101:47-56.
11. Lisman T*, et al.* (2006) A single high-affinity binding site for von Willebrand factor in collagen III, identified using synthetic triple-helical peptides. *Blood* 108:3753-3756.
12. Konitsiotis AD*, et al.* (2008) Characterization of high affinity binding motifs for the discoidin domain receptor DDR2 in collagen. *J Biol Chem* 283:6861-6868.
13. Giudici C*, et al.* (2008) Mapping of SPARC/BM-40/osteonectin-binding sites on fibrillar collagens. *J Biol Chem* 283:19551-19560.
14. Bradshaw AD, Sage EH (2001) SPARC, a matricellular protein that functions in cellular differentiation and tissue response to injury. *J Clin Invest* 107:1049-1054.
15. Rentz TJ*, et al.* (2007) SPARC regulates processing of procollagen I and collagen fibrillogenesis in dermal fibroblasts. *J Biol Chem* 282:22062-22071.
16. Fitzgerald MC, Schwarzbauer JE (1998) Importance of the basement membrane protein SPARC for viability and fertility in *Caenorhabditis elegans*. *Curr Biol* 8:1285-1288.
17. Martinek N, Shahab J, Saathoff M, Ringuette M (2008) Haemocyte-derived SPARC is required for collagen-IV-dependent stability of basal laminae in Drosophila embryos. *J Cell Sci* 121:1671-1680.
18. Hohenester E*, et al.* (1996) Structure of a novel extracellular Ca^{2+}-binding module in BM-40. *Nat Struct Biol* 3:67-73.

19. Hohenester E, Maurer P, Timpl R (1997) Crystal structure of a pair of follistatin-like and EF-hand calcium-binding domains in BM-40. *EMBO J* 16:3778-3786.
20. Sasaki T, Hohenester E, Göhring W, Timpl R (1998) Crystal structure and mapping by site-directed mutagenesis of the collagen-binding epitope of an activated form of BM-40/SPARC/osteonectin. *EMBO J* 17:1625-1634.
21. Bella J, Eaton M, Brodsky B, Berman HM (1994) Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science* 266:75-81.
22. Okuyama K*, et al.* (2006) Helical twists of collagen model peptides. *Biopolymers* 84:421-432.
23. Brandl M*, et al.* (2001) C-H···π-interactions in proteins. *J Mol Biol* 307:357-377.
24. Xie RL, Long GL (1995) Role of N-linked glycosylation in human osteonectin. Effect of carbohydrate removal by N-glycanase and site-directed mutagenesis on structure and binding of type V collagen. *J Biol Chem* 270:23212-23217.
25. Kaufmann B*, et al.* (2004) Structural variability of BM-40/SPARC/osteonectin glycosylation: implications for collagen affinity. *Glycobiology* 14:609-619.
26. Orgel JP, Irving TC, Miller A, Wess TJ (2006) Microfibrillar structure of type I collagen in situ. *Proc Natl Acad Sci USA* 103:9001-9005.
27. Maurer P*, et al.* (1995) The C-terminal portion of BM-40 (SPARC/osteonectin) is an autonomously folding and crystallisable domain that binds calcium and collagen IV. *J Mol Biol* 253:347-357.
28. Horii K, Kahn ML, Herr AB (2006) Structural basis for platelet collagen responses by the immune-type receptor glycoprotein VI. *Blood* 108:936-942.
29. Ichikawa O*, et al.* (2007) Structural basis of the collagen-binding mode of discoidin domain receptor 2. *EMBO J* 26:4168-4176.
30. Bella J, Brodsky B, Berman HM (1995) Hydration structure of a collagen peptide. *Structure* 3:893-906.
31. Maurer P*, et al.* (1997) Structural and functional characterization of the extracellular calcium-binding protein BM-40/secreted protein, acidic, rich in cysteine/osteonectin from the nematode *Caenorhabditis elegans*. *Eur J Biochem* 248:209-216.
32. Khoshnoodi J, Pedchenko V, Hudson BG (2008) Mammalian collagen IV. *Microsc Res Tech* 71:357-370.
33. Golbik R, Eble JA, Ries A, Kühn K (2000) The spatial orientation of the essential amino acid residues arginine and aspartate within the α1β1 integrin recognition site of collagen IV has been resolved using fluorescence resonance energy transfer. *J Mol Biol* 297:501-509.
34. Mayer U*, et al.* (1991) Calcium-dependent binding of basement membrane protein BM-40 (osteonectin, SPARC) to basement membrane collagen type IV. *Eur J Biochem* 198:141-150.
35. Bienkowska J*, et al.* (1997) The von Willebrand factor A3 domain does not contain a metal ion-dependent adhesion site motif. *J Biol Chem* 272:25162-25167.
36. Huizinga EG*, et al.* (1997) Crystal structure of the A3 domain of human von Willebrand factor: implications for collagen binding. *Structure* 5:1147-1156.
37. Nishida N*, et al.* (2003) Collagen-binding mode of vWF-A3 domain determined by a transferred cross-saturation experiment. *Nat Struct Biol* 10:53-58.

38. Romijn RA*, et al.* (2003) Mapping the collagen-binding site in the von Willebrand factor-A3 domain. *J Biol Chem* 278:15035-15039.
39. Hellings M*, et al.* (2004) Experimental indication for the existence of multiple Trp rotamers in von Willebrand Factor A3 domain. *Proteins* 57:596-601.
40. Jarvis GE*, et al.* (2008) Identification of a major GpVI-binding locus in human type III collagen. *Blood* 111:4986-4996.
41. Zong Y*, et al.* (2005) A 'Collagen Hug' model for *Staphylococcus aureus* CNA binding to collagen. *EMBO J* 24:4224-4236.

Figure Legends

Fig. 1. Crystal structure of SPARC FS-EC ΔαC bound to a 33-residue collagen peptide (stereo view). The FS and EC domains of SPARC are in green and cyan, respectively. Disulphide bridges are in pale pink, the glycan attached to N99 is in grey, and a calcium ion is shown as a purple sphere. The collagen peptide is shown as a Cα ribbon (leading chain, yellow; middle chain, orange; trailing chain, red). The chain termini, selected helices and the location of the αC deletion are labelled.

Fig. 2. Details of the SPARC-collagen interaction. (*A*) Sequence of the collagen peptide, indicating the one-residue stagger between the chains. The sequence numbering of the leading strand is indicated at the top, and the SPARC-binding motifs of the middle and trailing chains are coloured. (*B*) Surface representation of the collagen binding site of SPARC. The FS and EC domains of SPARC are in green and cyan, respectively. The collagen peptide is shown as a Cα ribbon (leading chain, yellow; middle chain, orange; trailing chain, red). Selected residues of the middle and trailing chains are shown in atomic detail and are labelled. (*C*) Interactions of SPARC with the trailing chain. SPARC is shown as a cartoon with semitransparent surface rendering. Selected SPARC residues are shown in atomic detail and are labelled. The collagen leading and middle chains are shown as a Cα ribbon and the trailing chain is shown in atomic detail. Hydrogen bonds are indicated by dashed lines. (*D*) Interactions of SPARC with the collagen middle chain.

Fig. 3. Structural changes within SPARC upon collagen binding. (*A*) Surface representation of the collagen binding region in apo SPARC. The FS and EC domains are in green and cyan, respectively. Residues implicated in collagen binding by mutagenesis (20) are in purple and labelled. W153 is shown in blue and the location of P244 (see text) is indicated. (*B*) Surface representation of the collagen-binding site in the SPARC-collagen complex. The colour scheme is the same as in *A*. The side chain of F23 of the collagen trailing chain is shown in atomic detail (red) to faciliate comparison with Figs. 1 and 2. (*C*) Stereo view of a superposition of the region surrounding the 'Phe pocket' in apo SPARC (light brown) and the SPARC-collagen complex (cyan). Selected residues are shown in atomic detail and are labelled. F23 of the collagen trailing chain is shown as a semitransparent space-filling model (red).

Fig. 4. Putative SPARC-binding sites in collagen IV. Shown are partial sequences of human collagen III (SwissProt entry P02461) and collagen IV (α1 chain, P02462; α2 chain, P08572). The SPARC-binding site in collagen III is highlighted; residues that are predicted to be strictly required for SPARC binding (see text) are in red, residues that should be apolar are in orange. The same colouring scheme is used to indicate the four putative SPARC-binding sites in collagen IV.

Supporting Information

Structural basis of sequence-specific collagen recognition by SPARC

Erhard Hohenester, Takako Sasaki, Camilla Giudici, Richard W. Farndale

and Hans Peter Bächinger

Materials and Methods

Peptide synthesis. The collagen peptide $(GPO)_3GPSGPRGQOGVMGFOGPKGPOGAO$ (O denotes 2*R*,4*R*-hydroxyproline) was synthesised with an ABI 433A synthesiser. Couplings were carried out on a Fmoc-PAL-PEG-PS resin (0.16 mmol/g) (Applied Biosystems). Fmoc-Gly-OH, Fmoc-Pro-OH, Fmoc-Ser(tBu)-OH, Fmoc-Arg(Pbf)-OH, Fmoc-Gln(Trt)-OH, Fmoc-Val-OH, Fmoc-Met-OH, Fmoc-Phe-OH, Fmoc-Lys(Boc)-OH and Fmoc-Ala-OH were from Applied Biosystems and Fmoc-4(*R*)Hyp(*t*Bu)-OH was from Novabiochem. O-(7-azabenzotriazol-1-yl)-1.1.3.3-tetramethyluronium hexafluorophosphate (HATU, 4.0 equivalents; Perseptive Biosystems) and diisopropylethylamine were used as the coupling reagent. The peptide was cleaved from the resin with Reagent R (trifluoroacetic acid:thioanisole-1,2:ethanedithiol:anisole, 90:5:3:2) and separated from the cleavage cocktail by precipitation with diethyl ether at 4 ^{o}C. The peptide was dissolved in 0.1% trifluoroacetic acid (TFA) and purified by preparative HPLC (Vydac® C18, 5μm, 300Å, 250 x 50 mm, W. R. Grace) using a 0-50% acetonitrile gradient in 0.1% TFA for elution. The peptide was characterised by amino acid analysis and electrospray/quadrupole/time-of-flight mass spectrometry (Q-tof micro, Waters Asociates). The correct amino acid composition was obtained and the mass was determined to be 3064.0 Da (calculated 3064.4 Da).

Protein production and complex formation. The human SPARC FS-EC ΔαC protein (residues 53-286 of SwissProt entry P09486, with residues 196-203 deleted) was expressed in HEK293 cells and purified as described (1). The SPARC-collagen complex

was formed by dissolving the lyophilised collagen peptide in a concentrated protein solution (molar peptide:protein ratio ~1.5:1) and purified by size exclusion chromatography on a 24 ml Superdex 200 column (GE Healthcare) with 0.01 M Na-HEPES pH 7.5, 0.14 M NaCl, 1 mM $CaCl_2$ as the running buffer. The complex eluted in a single sharp peak at 14.8 ml, compared with 15.7 ml for the free protein.

Crystallisation and data collection. The SPARC-collagen complex was concentrated to 15 mg/ml and crystallised at room temperature by the hanging drop method, using 0.1 M bis Tris-HCl pH 5.5, 0.25 M ammonium sulphate, 12-14% (w/v) PEG3350 as the precipitant. Large tetragonal bipyramids grew within two days and were flash-frozen in liquid nitrogen using 20% glycerol as cryoprotectant. A di-μ-iodo-bis(ethylenediamine)diplatinum (PIP) derivative was prepared by soaking a crystal in a 10% saturated PIP solution for 18 hours. Diffraction data were collected at 100 K on station 10.1 at the SRS Daresbury and processed with MOSFLM (www.mrc-lmb.cam.ac.uk/harry/mosflm) and programs of the CCP4 suite (2). Data collection statistics are listed in Table S1.

Structure solution and refinement. The SPARC-collagen structure was solved by a combination of molecular replacement and single isomorphous replacement with anomalous scattering (SIRAS). The SPARC EC ΔαC domain (residues 138-195/204-286) (1) could be located with PHASER (3, 4) with a translation function Z-score of 9.9, but all attempts to locate the FS domain or the collagen peptide were unsuccessful. Useful SIRAS phases were obtained from a poor PIP derivative with AutoSHARP

(Globalphasing Ltd, Cambridge, UK). The electron density map after density modification confirmed the position of the EC domain and revealed the position of a large part (residues 84-137) of the FS domain. Refinement of the incomplete SPARC model against the native 1 data with CNS (5) resulted in a R-factor of 0.451 (R_{free} 0.497) at 3.2 Å resolution and yielded an electron density map that showed strong density for the collagen peptide. Distinctive electron density for the amino acid side chains of the GVMGFO motif allowed the direction and register of the three collagen chains to be determined (Fig. S1). Multiple rounds of rebuilding and refinement with CNS lowered the R-factor to 0.284 (R_{free} 0.336). A final round of TLS refinement (four TLS bodies) with PHENIX (6) resulted in the final R-factor of 0.254 (R_{free} 0.320). The final model comprises SPARC residues 54-187 and 206-285, two *N*-acetyl glucosamine moities attached to N99, one calcium and one sulphate ion, as well as all 3x33 collagen residues. Residual difference electron density in the distal portion of the SPARC FS domain indicates that this part of the structure (which is distant from the SPARC-collagen interface) suffers from disorder, explaining the relatively high R-factors. Similar disorder was encountered in the structure determination of apo SPARC FS-EC (7). Phasing and refinement statistics are listed in Table S1.

Figures and analysis. The figures were made with PYMOL (www.pymol.org). Accessible surfaces were calculated with the CCP4 program AREAIMOL (2). Collagen superhelix parameters were calculated with a CNS script kindly provided by Jordi Bella (University of Manchester, UK).

References

1. Sasaki T, Hohenester E, Göhring W, Timpl R (1998) Crystal structure and mapping by site-directed mutagenesis of the collagen-binding epitope of an activated form of BM-40/SPARC/osteonectin. *EMBO J* 17:1625-1634.
2. CCP4 (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* 50:760-763.
3. McCoy AJ, Grosse-Kunstleve RW, Storoni LC, Read RJ (2005) Likelihood-enhanced fast translation functions. *Acta Crystallogr D Biol Crystallogr* 61:458-464.
4. Storoni LC, McCoy AJ, Read RJ (2004) Likelihood-enhanced fast rotation functions. *Acta Crystallogr D Biol Crystallogr* 60:432-438.
5. Brünger AT*, et al.* (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* 54:905-921.
6. Adams PD*, et al.* (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* 58:1948-1954.
7. Hohenester E, Maurer P, Timpl R (1997) Crystal structure of a pair of follistatin-like and EF-hand calcium-binding domains in BM-40. *EMBO J* 16:3778-3786.
8. Laskowski RA, MacArthur MW, Moss DA, Thornton JM (1996) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 26:283-291.
9. Brandl M*, et al.* (2001) C-H···π-interactions in proteins. *J Mol Biol* 307:357-377.

Figure Legends

Fig. S1. Stereo view of the electron density for the collagen peptide. Shown is a 3000K simulated annealing F_{obs}-F_{calc} omit map contoured at 2.0 σ. Residues 19-27 (GVMGFOGPK) of all three collagen chains were deleted from the coordinate file prior to refinement and map calculation. Superimposed on the map is the final refined collagen model (yellow, leading chain; orange, middle chain; red, trailing chain). Selected residues are labelled.

Fig. S2. Comparison of apo SPARC and the SPARC-collagen complex. (*A*). Superposition of apo SPARC (light brown) and SPARC in the collagen complex (cyan). The two structures were superimposed on 44 Cα atoms of EF-hand pair (r.m.s. deviation 0.59 Å) and are shown as Cα traces. The first and last residues of the polypeptide chains are labelled, as the EF-hands and selected helices. The FS domain (on the left) is positioned differently in the two structures, most likely as a result of crystal packing forces. (*B*) Calcium coordination in apo SPARC FS-EC ΔαC (1). Calcium ions are shown as pink spheres. Calcium-ligand bonds are indicated by solid black lines. Selected residues are shown in atomic detail and are labelled. (*C*) Calcium coordination in SPARC FS-EC ΔαC with bound collagen peptide (this study). Hydrogen bonds at the empty EF-hand 2 are indicated by dashed black lines.

Fig. S3. Sequence alignment of selected SPARC EC domains. The sequences are of human SPARC (SwissProt P09486), human SPARC-like/hevin (Q14515), *Xenopus*

laevis SPARC (P36378), *Drosophila melanogaster* SPARC (O97365) and *Caenorhabditis elegans* SPARC (P34714). Conserved residues are highlighted in yellow and cysteines are in red. Residues involved in calcium and collagen binding to human SPARC are indicated by pink circles and black inverted triangles, respectively. The sequence numbering and α-helices A-G of human SPARC are indicated above the alignment.

Table S1. Crystallographic statistics

	Native 1	Native 2	PIP derivative
Data Collection			
Space group	$P4_22_12$	$P4_22_12$	$P4_22_12$
Unit cell dimensions (Å)	a = b = 89.79	a = b = 90.19	a = b = 90.40
	c = 127.10	c = 127.12	c = 127.69
Beamline	SRS 10.1	SRS 10.1	SRS 10.1
Wavelength (Å)	0.980	0.980	1.045
Resolution (Å)	20-3.2 (3.37-3.20)[a]	20-3.5 (3.69-3.50)	20-3.8 (4.01-3.80)
Unique reflections	9046	7016	5589
Multiplicity	8.7 (8.3)	6.2 (6.1)	7.9 (8.2)
Completeness (%)	99.6 (100)	99.4 (100)	99.3 (100)
R_{merge}	0.085 (0.272)	0.073 (0.220)	0.118 (0.474)
Heavy Atom Phasing			
R_{deriv} with native 2			0.203
Heavy atom sites			1
Phasing power, acentric			0.71
Refinement			
Reflections	9019		
Atoms	2300		
R_{cryst}/R_{free}	0.254/0.320		
Rmsd bonds (Å)	0.009		
Rmsd angles (°)	1.4		
Average/Wilson B ($Å^2$)	87.3/74.0		
Ramachandran plot (%)[b]	70.0/29.1/0.9/0.0		

[a]Values in parantheses refer to the highest resolution shell.
[b]Percentage of residues in most favoured, additionally allowed, generously allowed, and disallowed regions (8).

Table S2. Collagen-SPARC contacts with d ≤ 4 Å

Collagen Atom	SPARC Atom	Distance (Å)	Contact Type
Leading Chain			
(V20 CG2	F92 CD2	4.51	closest contact with SPARC)
Middle Chain			
M21 SD	L242 CD1	3.27	vdW[a]
M21 CE	W153 CZ2	3.47	vdW
	W153 CH2	3.65	vdW
	L242 CD1	3.35	vdW
F23 CD2	M245 SD	3.75	vdW
	M245 CE	3.72	vdW
F23 CE2	M245 CE	3.50	vdW
O24 OD	M245 O	2.68	H-bond[b]
	E246 OE2	2.31	H-bond
Trailing Chain			
Q17 OE1	R184 CB	3.96	vdW
	R184 CG	3.78	vdW
	R184 CD	3.91	vdW
O18 O	R184 NH1	2.73	H-bond
G19 O	R184 NH2	2.87	H-bond
V20 CG1	V157 CG2	3.62	vdW
	T160 OG1	3.93	vdW
M21 O	W153 NE1	2.87	H-bond
F23 CD1	W153 CG	3.37	vdW
	W153 CD1	3.79	vdW
	W153 CD2	3.52	vdW
F23 CE1	W153 CD2	3.11	C-H···π H-bond?[c]
	W153 CE3	3.10	C-H···π H-bond?
	W153 CE2	3.80	vdW
	W153 CZ3	3.76	vdW
F23 CZ	M150 CA	3.88	vdW
	M150 CB	3.77	vdW
	M150 CG	3.65	vdW
F23 CD2	R149 NH2	3.68	vdW
	E246 OE2	3.14	C-H···π H-bond?
	E246 CD	3.22	C-H···π H-bond?
	E246 CG	3.68	vdW
	E246 OE1	3.66	vdW
F23 CE2	E246 CG	3.75	vdW
	F146 CE1	3.98	vdW
	F146 CZ	3.85	vdW
O24 O	R149 NE	3.09	H-bond
	R149 NH2	3.45	vdW
O24 OD	D152 CB	3.84	vdW
	D156 ND2	3.79	water-mediated H-bond?
K27 N	E145 OE2	2.87	H-bond

[a]van der Waals contact
[b]polar hydrogen bond
[c]hydrogen bond between C-H donor and π-electron acceptor (9)

Figure 1

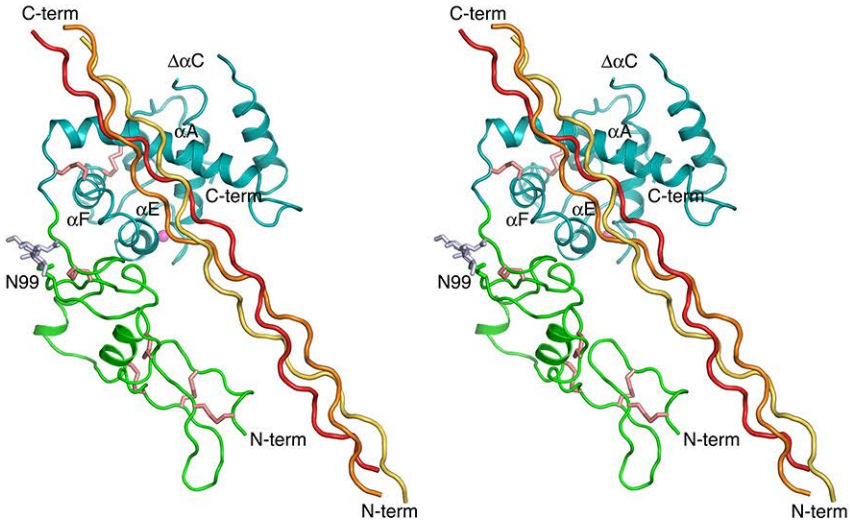
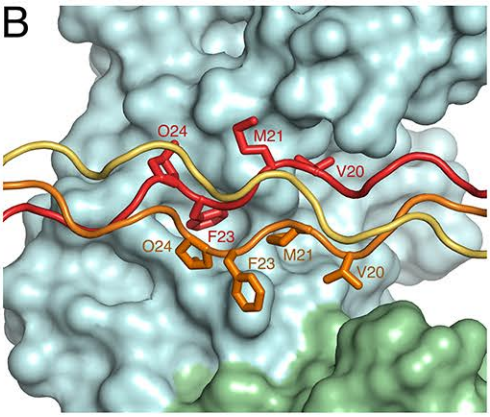

C-term
ΔαC
αA
C-term
αF
αE
N99
N-term
N-term
C-term
ΔαC
αA
C-term
αF
αE
N99
N-term
N-term

Figure 2

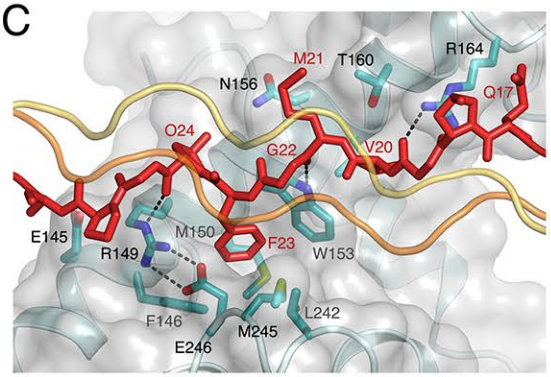
A

```
                    10          20          30
Leading   GPOGPOGPOGPSGPRGQOGVMGFOGPKGPOGAO
Middle     GPOGPOGPOGPSGPRGQOGVMGFOGPKGPOGAO
Trailing    GPOGPOGPOGPSGPRGQOGVMGFOGPKGPOGAO
```

B



C



D

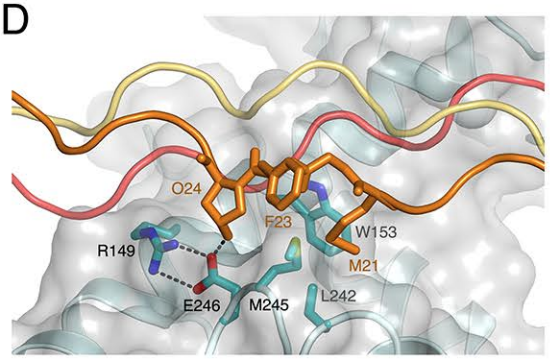


Figure 3

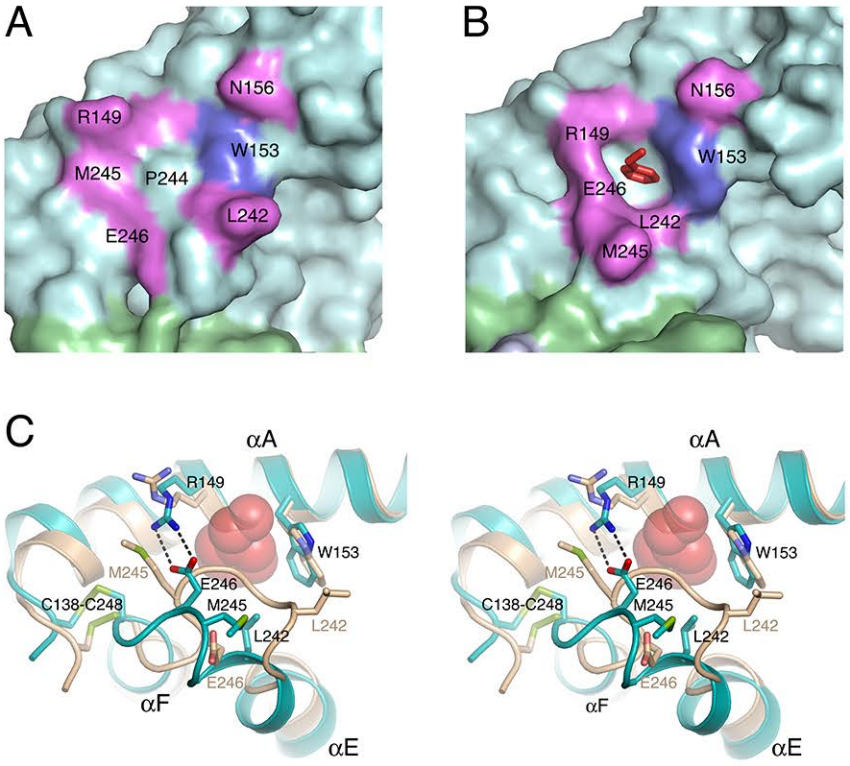

A
N156
R149
W153
M245
P244
L242
E246
B
N156
R149
W153
E246
L242
M245
C
αA
R149
W153
M245
E246
C138-C248
M245
L242
L242
E246
αF
αE
αA
R149
W153
M245
E246
C138-C248
M245
L242
L242
E246
αF
αE

Figure 4

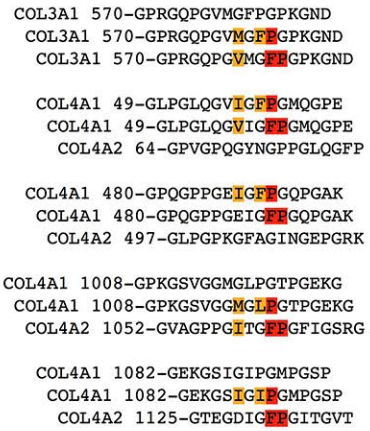

COL3A1 570-GPRGQPGVMGFPGPKGND
COL3A1 570-GPRGQPGVMGFPGPKGND
COL3A1 570-GPRGQPGVMGFPGPKGND
COL4A1 49-GLPGLQGVIGFPGMQGPE
COL4A1 49-GLPGLQGVIGFPGMQGPE
COL4A2 64-GPVGPQGYNGPPGLQGFP
COL4A1 480-GPQGPPGEIGFPGQPGAK
COL4A1 480-GPQGPPGEIGFPGQPGAK
COL4A2 497-GLPGPKGFAGINGEPGRK
COL4A1 1008-GPKGSVGGMGLPGTPGEKG
COL4A1 1008-GPKGSVGGMGLPGTPGEKG
COL4A2 1052-GVAGPPGITGFPGFIGSRG
COL4A1 1082-GEKGSIGIPGMPGSP
COL4A1 1082-GEKGSIGIPGMPGSP
COL4A2 1125-GTEGDIGFPGITGVT

Figure S1

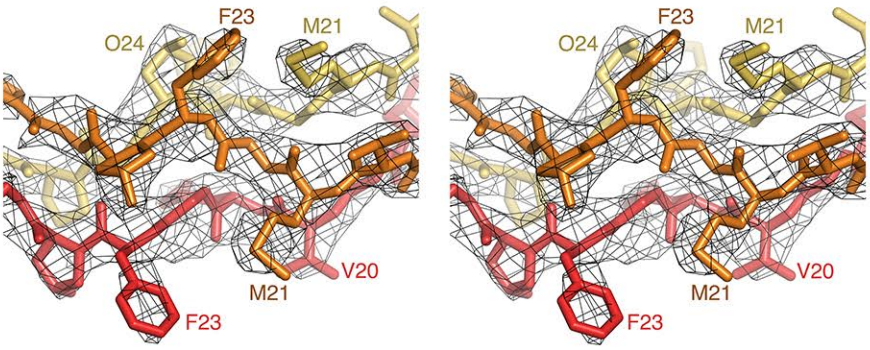


Figure S2

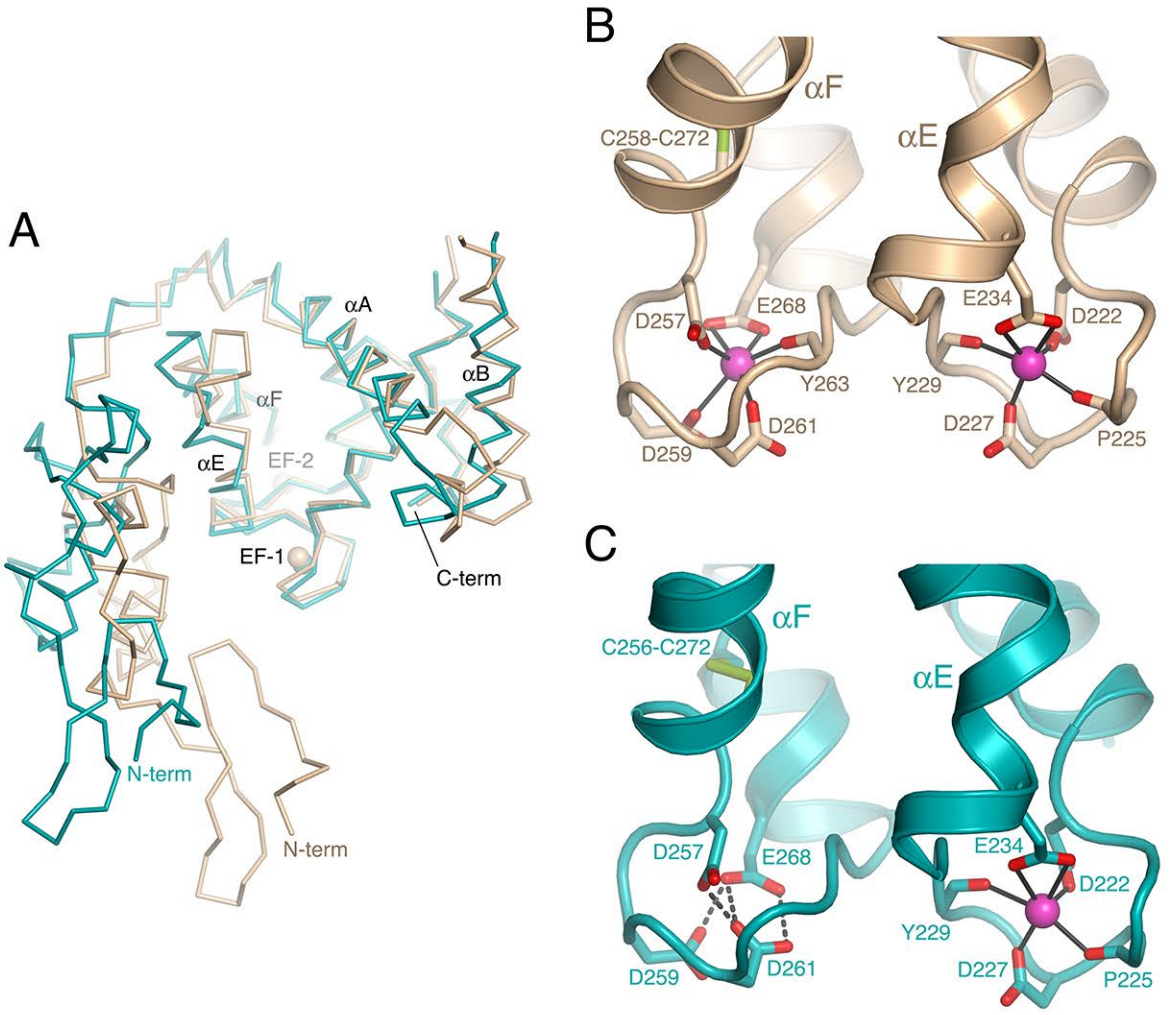
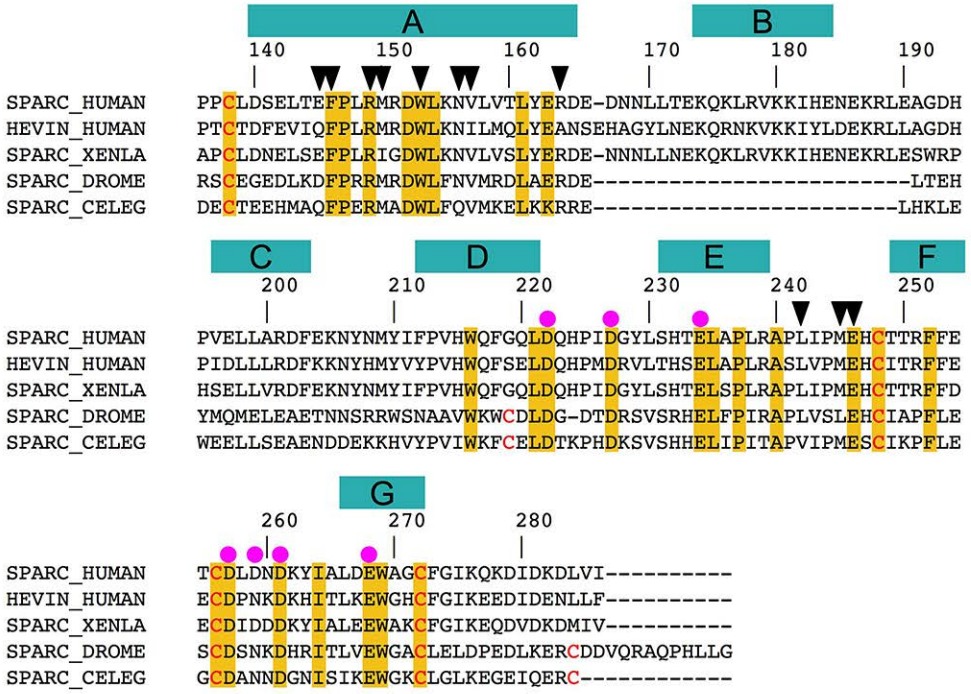

A
αA
αB
αF
αE
EF-2
EF-1
C-term
N-term
N-term
B
αF
αE
C258-C272
D257
E268
Y263
D261
D259
E234
D222
Y229
D227
P225
C
αF
αE
C256-C272
D257
E268
D259
D261
E234
D222
Y229
D227
P225

Figure S3


A
B
C
D
E
F
G
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
SPARC_HUMAN PPCLDSELTEFPLRMRDWLKNVLVTLYERDE-DNNLLTEKQKLRVKKIHENEKRLEAGDH
HEVIN_HUMAN PTCTDFEVIQFPLRMRDWLKNILMQLYEANSEHAGYLNEKQRNKVKKIYLDEKRLLAGDH
SPARC_XENLA APCLDNELSEFPLRIGDWLKNVLVSLYERDE-NNNLLNEKQKLRVKKIHENEKRLESWRP
SPARC_DROME RSCEGEDLKDFPRRMRDWLFNVMRDLAERDE-----------------------LTEH
SPARC_CELEG DECTEEHMAQFPERMADWLFQVMKELKKRRE-----------------------LHKLE
SPARC_HUMAN PVELLARDFEKNYNMYIFPVHWQFGQLDQHPIDGYLSHTELAPLRAPLIPMEHCTTRFFE
HEVIN_HUMAN PIDLLLRDFKKNYHMYVYPVHWQFSELDQHPMDRVLTHSELAPLRASLVPMEHCITRFFE
SPARC_XENLA HSELLVRDFEKNYNMYIFPVHWQFGQLDQHPIDGYLSHTELSPLRAPLIPMEHCTTRFFD
SPARC_DROME YMQMELEAETNNSRRWSNAAVWKWCDLDG-DTDRSVSRHELFPIRAPLVSLEHCIAPFLE
SPARC_CELEG WEELLSEAENDDEKKHVYPVIWKFCELDTKPHDKSVSHHELIPITAPVIPMESCIKPFLE
SPARC_HUMAN TCDLDNDKYIALDEWAGCFGIKQKDIDKDLVI----------
HEVIN_HUMAN ECDPNKDKHITLKEWGHCFGIKEEDIDENLLF----------
SPARC_XENLA ECDIDDDKYIALEEWAKCFGIKEQDVDKDMIV----------
SPARC_DROME SCDSNKDHRITLVEWGACLELDPEDLKERCDDVQRAQPHLLG
SPARC_CELEG GCDANNDGNISIKEWGKCLGLKEGEIQERC------------