Back, C., Higman, V. A., Le Vay, K., Patel, V. V., Parnell, A. E., Frankel, D., ... Race, P. R. (2020). The streptococcal multidomain fibrillar adhesin CshA has an elongated polymeric architecture. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.RA119.011719

The streptococcal multidomain fibrillar adhesin CshA has an elongated polymeric architecture

Catherine R. Back[‡§¶], Victoria A. Higman[**], Kristian LeVay[§†], Viren V. Patel[§], Alice E. Parnell[§¶], Daniel Frankel[‖], Howard F. Jenkinson[‡], Steven G. Burston[§¶], Matthew P. Crump[¶], Angela H. Nobbs[‡1], Paul R. Race[§¶2]**

From [‡]Bristol Dental School, University of Bristol, Lower Maudlin St, Bristol, BS1 2LY, UK; [§]School of Biochemistry, University of Bristol, University Walk, Bristol, BS8 1TD, UK; [¶]BrisSynBio Synthetic Biology Research Centre, Life Sciences Building, Tyndall Avenue University of Bristol, BS8 1TQ, UK; [**]School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, UK; [†]Bristol Centre for Functional Nanomaterials, HH Wills Physics Laboratory, University of Bristol, BS8 1TL, UK; [‖]School of Engineering, Newcastle University, Newcastle-upon-Tyne, NE1 7RU, UK

Running title: Structure of *Streptococcus gordonii* CshA

* To whom correspondence should be addressed: [1]Dr. Angela H. Nobbs, School of Oral & Dental Sciences, University of Bristol, Lower Maudlin Street, Bristol, BS1 2LY, UK, Telephone:+44 (0)1173424779, Email: Angela.Nobbs@bristol.ac.uk. [2]Dr. Paul R. Race, School of Biochemistry, University of Bristol, Biomedical Sciences Building, University Walk, Bristol BS8 1TD, UK, Telephone: +44 (0)1173311835, Email: Paul.Race@bristol.ac.uk.

Keywords: adhesin, fibril, protein folding, microbiology, bacterial pathogenicity, NMR spectroscopy, small-angle X-ray scattering (SAXS), virulence factor, biofilm, CshA

ABSTRACT

The cell surfaces of many bacteria carry filamentous polypeptides termed adhesins that enable binding to both biotic and abiotic surfaces. Surface adherence is facilitated by the exquisite selectivity of the adhesins for their cognate ligands or receptors and is a key step in niche or host colonization and pathogenicity. *Streptococcus gordonii* is a primary colonizer of the human oral cavity and an opportunistic pathogen as well as a leading cause of infective endocarditis in humans. The fibrillar adhesin CshA is an important determinant of *S. gordonii* adherence, forming peritrichous fibrils on its surface that bind host cells and other microorganisms. CshA possesses a distinctive multidomain architecture comprising an N-terminal target-binding region fused to seventeen ~100-amino-acid-long repeat domains (RDs). Here, using structural and biophysical methods, we demonstrate that the intact CshA repeat region (CshA_RD1-17, domains 1–17) forms an extended polymeric monomer in solution. We recombinantly produced a subset of CshA RDs and found that they differ in stability and unfolding behavior. The NMR structure of CshA_RD13 revealed a hitherto unreported all β-fold, flanked by disordered interdomain linkers. These findings, in tandem with complementary hydrodynamic studies of CshA_RD1-17, indicate that this polypeptide possesses a highly unusual dynamic transitory structure characterized by alternating regions of order and disorder. This architecture provides flexibility for the adhesive tip of the CshA fibril to maintain bacterial attachment that withstands shear forces within the human host. It may also help mitigate deleterious folding events between neighboring RDs that share significant structural identity without compromising mechanical stability.

Bacteria occupy almost every ecological niche on Earth (1,2). Their capacity to colonize diverse environments is in part enabled by their

ability to adhere to the surfaces of materials and other cells. Adherence allows anchorage and persistence within a defined environment, confers significant evolutionary advantage, and promotes bacterial infection in animals and humans (3,4). Identifying and characterizing the cellular machineries employed by bacteria to adhere and colonize is of broad fundamental interest and may inform the development of anti-infective agents, medical devices or vaccines (5,6).

Frequently, bacteria utilize proteinaceous surface decorations termed adhesins to facilitate attachment to extracellular target molecules. Different adhesins recognize and bind different (a)biotic targets, and there is considerable diversity in the molecular architectures of these important polypeptides. Larger filamentous adhesins may be grouped into one of two categories based on their distinguishing structural features; these are pili and fibrils. Pili have been implicated in numerous physiological processes and are found in both Gram-positive and Gram-negative bacteria (7-9). Fibrillar adhesins are produced by a wide variety of bacteria. They exhibit considerable sequence diversity and much still remains to be learned about their structures and functions. Fibrils are usually composed of a single polypeptide, which is covalently anchored to the cell wall *via* a C-terminal LPxTG motif (10-12).

Streptococcus species, including both commensal strains and pathogens, are prodigious producers of fibrillar adhesins (13-16). *Streptococcus gordonii*, a pioneer oral bacterium and opportunistic pathogen, employs the fibrillar adhesin Cell surface hydrophobicity protein A (CshA) to enable binding to host cell surfaces and other microorganisms (17). This ~259 kDa polypeptide shares <10% sequence identity to any protein of known structure (17-19). CshA possesses a distinctive multidomain architecture, comprising an N-terminal signal peptide (41 aa residues), a non-repetitive target binding region (778 aa), a repetitive region composed of 17 sequentially-arrayed repeat domains (RDs; ~100 aa each), and an LPxTG anchor (Figure 1). CshA forms peritrichous fibrils of ~60 nm on the surface of *S. gordonii* (17) and heterologous expression of this protein on the surface of *Enterococcus faecalis* results in the formation of a dense furry layer comprised of multiple closely associated CshA polypeptides, which confers adhesive properties (17). Similarly, Δ*cshA* strains of *S. gordonii* show reduced binding to other oral microorganisms and host molecules, including fibronectin (Fn) (18-20). Recently, the molecular details of host Fn binding by CshA were established, with this polypeptide shown to bind Fn *via* a distinctive 'catch-clamp' mechanism, mediated by discrete domains within the non-repeat region of the protein (21). This mode of binding involves the action of the intrinsically disordered N-terminal domain of the protein and its neighboring ligand-binding domain, which function in concert to form a robust protein-protein interaction *via* a readily dissociable pre-complex intermediate.

In this study, using a combination of structural and biophysical methods, we show that the > 175 kDa multi-domain repeat region of CshA (CshA_RD1-17) adopts an elongated polymeric structure in solution, with a distinctive conformation dictated by the interplay of fully and partially ordered domains, and intrinsically disordered regions. Equilibrium folding studies of individual CshA repeat domains reveal diversity in the stabilities and unfolding profiles of these proteins, despite their often considerable (>90%) sequence identities. The NMR structure of CshA_RD13 has been determined, which identifies a previously unreported all beta-fold flanked on either terminus by unstructured linker regions. Complementary AUC and SAXS studies of CshA_RD1-17 provide support for the CshA repeat region adopting a transitory structure characterized by alternating regions of order and disorder. Together, our data suggest a molecular architecture within which individual repeat domains contribute additive strength to the intact polypeptide, but also minimizes the likelihood of domain misfolding that may arise as a consequence of high sequence and structural identity to adjacent RDs. This is enabled *via* the acquisition of destabilizing mutations that preclude the adoption of a fully folded state. Our work identifies a distinctive polymeric protein architecture and resolves the molecular intricacies of its structure and organization. In turn, this provides greater insight regarding the capacity for bacterial adhesins to promote colonization of sites within the host that are continuously exposed to the flow of blood, saliva or tissue fluids.

RESULTS

The intact CshA repeat region adopts an extended polymeric structure in solution

Consistent with previous domain assignments, the repeat region of CshA was considered to comprise residues 820-2500 of the 2507 amino acid full-length CshA polypeptide (21) (Figure 1). The intact CshA repeat region, from herein referred to as CshA_RD1-17, was amplified from *S. gordonii* DL1 (22) chromosomal DNA and cloned into the pOPINF expression vector (23) (Table S1). The resulting construct was used to facilitate over-expression of an N-terminally hexa-histidine tagged variant of CshA_RD1-17 in *Escherichia coli*, and the resulting recombinant material was purified to homogeneity using a two-step process. CshA_RD1-17 was found to be a homogeneous, monodisperse species in solution, of >95% purity. Analysis of CshA_RD1-17 using circular dichroism (CD) spectroscopy, followed by deconvolution of the resulting spectrum into secondary structural elements, revealed the protein to be predominantly beta sheet (~45%), with a significant disorder content (~39%; Figure 2*A*). Sedimentation velocity analytical ultracentrifugation (SV-AUC) confirmed that the polypeptide is monomeric and adopts an extended configuration in solution with an $f/f_0 = 2.84$ (Figure 2*B* and Table S2). Complementary small angle X-ray scattering (SAXS) analysis (Figure 2*C* and Table S3) provided further evidence that CshA adopts an elongated structure, with a radius of gyration of 120 Å and maximum diameter of 408 Å, as derived from the pair distance distribution (*P(r)*) function (Figure 2*D*). Structural disorder is apparent from the Kratky plot, which diverges from the baseline at high *q*, and the Porod exponent, which is lower than observed for a well folded globular protein (Figure 2*E*). The structural disorder evident from CD and SAXS analysis suggests a flexible dynamic structure, in keeping with the biological role of CshA. The measured scattering data are well described by the flexible cylinder model (Figure 2*F* and Table S4), in which CshA is characterized by a higher Kuhn length and lower contour length than that expected for a random coil. The large deviation from random coil behaviour is consistent with a significant proportion of folded regions in the solution structure. These data imply that the polypeptide adopts an elongated, flexible ultrastructure in solution that occupies an ensemble of configurations.

Individual CshA repeat domains exhibit varying stabilities and unfolding behaviours

Having established the solution ultrastructure of CshA_RD1-17, we next sought to investigate the molecular origins of the polypeptide's physical properties. Comparative sequence analysis of assigned CshA repeat domains reveals considerable variation in the amino acid sequences of these regions (Figs. 3*A* and S1). The repeat region comprises a central core of domains with very high sequence identity (domains 3-14) punctuated by the deviant repeat domain 7. The sequence of this domain diverges significantly from those of the other sixteen repeat domains that comprise the intact repeat region. Surprisingly, a significant number of adjacent domains located within the central 3-14 core exhibit high sequence identity. Domains 3 and 4; 5 and 6; 10 and 11; 11 and 12; and, 12 and 13; share >90% sequence identity (Figure 3*A*), an arrangement that contravenes current dogma regarding the organization of tandemly-arrayed domains within multi-domain proteins (24). The sequence identities of the terminal domains of CshA_RD1-17, namely 1, 15, 16 and 17, are significantly lower than those identified in the central core region. Interestingly, in addition to repeat domain 17, domains 6, 11, 12, and 13 all possess a C-terminal LPxTG cell-wall anchor motif, suggesting that evolutionary pressure to present the adhesive non-repeat region of CshA at a maximal distance from the cell surface may have driven extension of the repeat region *via* gene duplication.

In an effort to explore the structural significance of sequence variation between individual CshA repeat domains a sub-set of these proteins were cloned, recombinantly over-expressed in *E. coli* and purified to homogeneity using the same general strategy (Table S1). Representative domains were selected covering a breadth of amino acid sequences. These were repeat domains 1, 3, 5, 7 and 13. Each could be readily produced in high quantities and to high purities (>95%, as judged by SDS-PAGE analysis). The stabilities and unfolding behaviours of each of these proteins was assessed *in vitro* by monitoring their unfolding in the presence of increasing concentrations of the chemical denaturant urea (Figs. 3*B* and Table 1). Unfolding behaviour was monitored by intrinsic tyrosine fluorescence, exploiting the presence of at least one such residue in each of the repeats 1, 3, 5,

7 and 13. Of the isolated domains examined, CshA_RD13 exhibited the highest overall stability (-3.42 kcal mol^{-1}) while remarkably, CshA_RD5 showed no fluorescence intensity change when titrated with urea, despite the 91% sequence identity with repeat 13, including the two tyrosine residues at precisely the same positions, 52 (residue Tyr2084) and 92 (residue Tyr2123) (Figure S1). CD spectroscopy of repeat domain 5 also indicated that this domain was largely unstructured, even in the absence of urea (data not shown). While CshA_RD3 and CshA_RD7 are less stable than CshA_RD13, they do exhibit a mildly cooperative unfolding transition, while CshA_RD1 is barely stable even in the absence of urea, but also exhibits a weakly cooperative unfolding transition. Complimentary size exclusion chromatography (SEC) analyses of individual CshA repeat domains provides further support for variability in the degree of foldedness of these proteins (Figure S2). The largely unfolded CshA_RD5 elutes earlier from an SEC column than its better folded counterparts, and significantly earlier than the well folded CshA_RD13.

Solution structure of CshA_RD13

In an effort to provide a structural framework for the observed biophysical properties of CshA_RD1-17, CshA_RD13, which possesses the highest cross-domain sequence identity to all other CshA repeat domains (Figure 3*A*), was selected for structure elucidation. Of the five single repeat domains produced recombinantly, CshA_RD13 has the greatest tolerance to urea unfolding, suggestive of high stability (Figure 3*B*). The structure of this protein was determined using solution NMR (Figs. 4, S3 and S4). Assignment proved challenging due to repetitive sequence motifs and a high degree of mobility leading to both the absence of some signals and the doubling (or more) of others (Figure 4*A*). Nonetheless, a high degree of assignment was achieved for the core region of the protein covering residues 2053-2130 (Table 2). The N-terminal region (2032-2052 plus a 19-residue tag) was found to be largely unstructured with few inter-residue NOEs and no unambiguously assignable long-range NOEs. For this reason, no structural restraints were included for this part of the sequence and the structure was only calculated and validated for residues 2053-2130. In addition to the high degree of disorder in the N-terminal part of CshA_RD13, several other regions of slow exchange (ms) were detected. Two sets of NMR signals were observed for the initial N-terminal loop comprised of residues 2053-2062, of which only the major set was used for structure calculations. A hydrogen-deuterium exchange experiment showed that the Val2059 NH group is involved in a hydrogen bond which persists for over an hour, suggesting that interconversion between these conformations is either very slow or, more likely, that they are very similar and both involve a hydrogen bond between Val2059H and Asp2056O (as determined from initial structure calculations conducted without hydrogen bond restraints). Multiple conformations were also observed for residues Asp2113 and Asn2115, which lie in the β5-β6 loop. The β5-β6 loop lies adjacent to the N-terminal loop suggesting slow exchange between these two regions may be coupled. The β4-β5 loop (Pro2096-Pro2106) and C-terminal tail (Ser2124-Val2130) are both ill-defined in the structural ensemble (Figure 4*B*), which is in part due to several broad, missing or unassigned signals and thus a low density of structural restraints that might reflect the underlying dynamics of these regions. Several residues along the outside edge of the β3 and β5 strands have NOEs that could not be assigned to residues within the globular domain. Most likely these arise from interactions with the N-terminal tail of the protein, though no unambiguous assignment to particular residues was possible.

CshA_RD13 adopts a β-sandwich fold comprising two 3-stranded anti-parallel β-sheets arranged at an angle of ~35° relative to one another (Figure 4*C*). The two sheets are connected by an 11 amino acid linker that fuses β4 to β5 and forms an extended loop that wraps around the C-terminal apex of the protein. The interface between the two β-sheets is predominantly hydrophobic and forms the compact core of the protein (Figure 4*D*). There is a high degree of amino acid sequence conservation in this region in other CshA repeat domains, suggesting that each domain retains this unique core fold. In addition to the highlighted hydrophobic residues, the two tyrosine residues of CshA_RD13 (Tyr2084 and Tyr2123) reside at the β-sandwich interface (Figure 4*D*), adding credence to the validity of our folding studies. Assessment of the solvation state of this pair of residues is likely to provide an accurate measure of protein unfolding. The overall shape of CshA_RD13 can be likened to a cylinder, which is tapered at both

termini. The terminal regions of the protein present sizeable patches of charge, suggestive that neighbouring repeat domains may be able to engage in complementary charge-charge interactions with one another. Four of the five residues universally conserved in all seventeen repeat domains (Gly2082, Gly2090, Gly2102 and Asp2113 in CshA_RD13) contribute to the constrainment of tight turns between individual β-strands (β2-β3, β3-β4, β4-β5 and β5-β6; SI). The 5th is located in the disordered N-terminal interdomain linker.

CshA_RD1-17 adopts a transitory dynamic structure comprising alternating regions of order and disorder

To reconcile our structural and biophysical data we attempted to construct a pseudo-atomic model describing the molecular architecture of CshA_RD1-17 in its entirety. The partial foldedness of the polypeptide implied from our SAXS and CD data, in addition to the variations in unfolding behaviour of selected repeat domains, and the observation of disordered linker regions at either terminus of the CshA_RD13 NMR structure, suggests that CshA may adopt a structure comprised of alternating regions of order and disorder. To verify this model, we applied an ensemble optimization method (EOM) to our SAXS data (Figure 5 and Table 3). Homology models of each repeat domain were generated and used to formulate pseudo-atomic models describing the molecular architecture of CshA_RD1-17, in which well folded domains are alternated with disordered regions approximated by a random coil. The data were well described by a model containing all 17 RD homology structures, although the calculated ensemble R_g (99 Å) was lower than that determined experimentally (120 Å) (Figure 5 and Table 3). The contour length of this structure is approximately 660 Å, more than half of that determined from the data using the flexible polymer model. However, this model underestimates the proportion of disordered structure as measured using CD. To compensate, only repeat domains predicted to be largely ordered (1, 3-4, 7-8, 14-16) were included in the model, yielding an ensemble with an average R_g in agreement with our experimentally measured value (Figure 5 and Table 3). These findings indicate that the ultrastructure of CshA_RD1-17 does not adhere to a standard 'beads-on-a-string' configuration, wherein individual well-folded RDs are arranged in a defined sequence within the polypeptide chain, but rather a highly dynamic architecture wherein a subset of RDs fail to adopt a fully folded state, thus leading to a highly dynamic transitory structure, dominated by the interplay of ordered, disordered and partially ordered regions.

DISCUSSION

Fibrillar adhesins are an important family of bacterial surface proteins that make significant contributions to environmental and host colonization, biofilm formation, host tissue invasion and pathogenicity. As virulence factors they represent attractive targets for the development of therapeutic strategies and interventions. Although many fibrillar adhesins have been identified in commensal and pathogenic bacteria, only a small number of these proteins have been subjected to detailed molecular level characterization. Examples include SasG, M protein, and the AgI/II family polypeptides (10,12,26-33). Each of these adhesins exploits a startlingly disparate molecular mechanism to facilitate the formation of fibrillar structures on the bacterial cell surface.

The *S. gordonii* fibrillar adhesin CshA plays an important role in host colonization. CshA possesses a distinctive modular architecture that comprises seventeen β-sandwich domains fused in series by flexible linkers. Although there is diversity in the sequences of individual repeat domains, amino acid sequence analysis suggests that each retains a conserved hydrophobic core that forms the basis of a compact protein fold. The structure of the representative repeat domain CshA_RD13 has been elucidated and provides a valuable test subject for understanding CshA repeat domain structure and function. The high degree of mobility in CshA_RD13 made assignment and structure calculation for this protein challenging; nonetheless, the core globular part of the protein is well defined (Figure 4). DALI analysis of CshA_RD13 failed to identify any closely related structural homologues of the protein and technically the domain exhibits a new fold. However, the flattened β-sandwich is reminiscent of Ig domains found in many other repeat domain containing proteins such as titin and cadherin. (25).

Folding studies of individual CshA RDs reveals remarkably variable stabilities considering their high sequence identities (Figure 3). Five of the

repeat domains (1, 3, 5, 7 and 13) were expressed individually and subjected to equilibrium unfolding to assess their relative stabilities. Repeat domains 3, 7 and 13 all displayed a cooperative unfolding transition with a relatively small free energy of folding, although not unusual for small domains (for example, see Gruszka *et al*., 2015 (27)). Equilibrium unfolding of CshA_RD1 revealed a weakly cooperative transition ($m_{D\text{-}N}$ = 0.66 kcal mol^{-1} M^{-1}) and only very marginal stability (0.64 kcal mol^{-1}), indicating that a significant proportion of the molecules are unfolded even in native conditions. Since CshA_RD1 is markedly divergent from all of the other repeat domains then it is difficult to relate differences in sequence to changes in stability. Interestingly, the sequences of the terminal repeat domains CshA_RD1 and CshA_RD17 differ considerably from those located centrally within the polypeptide. This may reflect the fact that they have coevolved to be adjacent to a non-repeat domain and the cell wall respectively.

As this repeat follows the non-repetitive region in the overall CshA structure then it may require the presence of that region to interact and stabilise it. No transition could be observed at all with CshA_RD5 which is surprising since it has 91% identity with CshA_RD13. An examination of the differences between the primary sequences of CshA_RD5 and CshA_RD13 with respect to the NMR structure of the latter suggest some differences that may be responsible for destabilizing CshA_RD5 relative to CshA_RD13. Thr2077 on β1b and both Pro2118 and Thr2122 on β5 in CshA_RD13 are all solvent exposed to some degree and have been substituted with valine, leucine and isoleucine respectively in CshA_RD5, leading to unfavourable exposure of hydrophobic residues to the aqueous solvent. Pro2079 which forms part of a Type II β-turn between strands β1b and β2 is substituted with a serine, which statistically has a greater preference for Type I β-turns.

Mapping amino acid conservation across all seventeen repeat domains onto the structure of CshA_RD13 indicates partial conservation of hydrophobic residues that reside within the hydrophobic cores of each RD. The central section of CshA_RD1-17 comprises 12/13 serially-arrayed repeat domains that possess a high degree of sequence identity and appear closely structurally related (Fig 3*A*). The sequential arrangement of high similarity domains is at odds with known sequence to folding relationships in tandemly-arrayed protein domains, where sequence disparity between neighboring domains is postulated to minimize protein misfolding (24). It is tempting to speculate that interdomain linker length and disorder plays an important role in this process, ensuring that the spatial distance between neighboring domains is sufficient to allow each individual domain to adopt its fully folded conformation prior to translation of its superseding neighbor. However, what is clear from our folding studies, and is corroborated by hydrodynamic analysis of CshA_RD1-17, is that a subset of CshA RDs do not adopt a well-folded conformation either alone, or in the context of the intact CshA polypeptide. This generalized loss in foldedness appears to arise due to the acquisition of destabilizing mutations within the hydrophobic core of some repeat domains. Significantly, these mutations appear to arise in instances where there is significant amino acid sequence identity to neighboring domains (Figs. 3*A* and S1). This may represent a strategy to minimize the likelihood of inter-domain misfolding events, thus mitigating adhesin aggregation on the bacterial cell surface. Alternatively, the solvent-exposed hydrophobic residues may help to mediate interaction between CshA polypeptides during assembly of the cell-surface adhesive layer.

The functional significance of the dynamic transitory structure of CshA_RD1-17 is yet to be unambiguously established, however, it is unquestionable that the combination of folded and partially folded regions will confer a high degree of flexibility to the polypeptide. This may enable the optimal projection of CshA's adhesive tip from the *S. gordonii* cell surface and in doing so maximize the capture radius of the adhesin. In addition, the partially folded structure may provide a mechanism of force damping following fibronectin binding. This could offer a mechanical advantage by mitigating the effects of sheer forces following target engagement. This would be of particular significance in the bloodstream, where it is necessary for *S. gordonii* to maintain an intimate association with the surface of host cells whilst resisting the force of blood flow. The transitory structure of CshRD1-17 would provide a deformable tether with the capacity to dissipate the kinetic energy of binding under flow.

In summary, here we report the identification and characterisation of an entirely new architecture for multi-domain bacterial surface proteins as typified by the *S. gordonii* adhesin CshA. This ultrastructure is characterized by the presence of fully- and partially-folded repeat domains, along with regions of intrinsic disorder, which affords a dynamic yet mechanically robust polymeric structure. Our study extends the diversity of natural protein architectures that are employed to enable microbial adherence to biotic and abiotic substrata, and provides new insight into the capacity for bacteria to adhere and persist at sites exposed to shear forces. Moreover, this information establishes a foundation for the development of interventions that target CshA and related polypeptides that can be applied to disease prevention and anti-biofouling strategies.

EXPERIMENTAL PROCEDURES

Gene cloning

DNA sequences encoding CshA_RD1-17, CshA_RD1, CshA_RD3, CshA_RD5, CshA_RD7 and CshA_RD13 were amplified from *Streptococcus gordonii* DL1 (22) chromosomal DNA using appropriate primers (Table S1), incorporating appropriate consensus sequences for subsequent cloning into the expression vector pOPINF (23), pre-cut with HindIII and KpnI. Ligations were performed using the In-Fusion™ (Clontech) cloning system as per the manufacturer's instructions. The resulting constructs encode N-terminally hexa-histidine tagged variants of each of the proteins under investigation. The sequences of all constructs were verified by DNA sequencing before being transformed into *Escherichia coli* BL21 (DE3) cells for protein expression.

Protein expression

For the expression of unlabelled CshA_RD 1-17, CshA_RD1, CshA_RD3, CshA_RD5, CshA_RD7 and CshA_RD13 cultures of *E. coli* BL21 (DE3) cells harbouring the respective expression plasmid were grown with shaking (200 rpm) in 1 L of LB (Luria-Bertani) broth, supplemented with carbenicillin (50 μg ml^{-1}), at 37 °C, to A_{600} = 0.4-0.6. Protein expression was induced by the addition of isopropyl β-galactopyranoside (IPTG) to a final concentration of 1 mM, and the cell cultures transferred to 20 °C, with shaking at 200 rpm, and grown for a further 16 h. For expression of ^{15}N labelled CshA_RD13, a culture (100 mL) of *E. coli* BL21 (DE3) cells harbouring CshA_RD13::pOPINF was grown overnight at 37 °C with shaking at 200 rpm. Cells were harvested by centrifugation, washed in resuspension buffer (50 mM Tris-HCl, 150 mM NaCl, pH 7.5), and used to inoculate 1 L of M9 minimal auto-induction media (50 mM KH_2PO_4, 25 mM Na_2HPO_4 (pH 6.8), 10 mM NaCl, 0.5% glycerol, 0.05% glucose, 0.2% α-lactose, 2 mM MgSO4) supplemented with carbenicillin (50 μg ml^{-1}), trace elements (10 mL, 100 x, to a final concentration of 13.4 mM EDTA, 3.1 mM $FeCl_3.6H_2O$, 0.62 mM $ZnCl_2$, 76 μM $CuCl_2.2H_2O$, 42 μM $CoCl_2.6H_2O$, 162 μM H_3BO_3, 8.1 μM $MnCl_2.6H_2O$) and 1 g/L $^{15}NH_4Cl$. Cells were grown with shaking at 37 °C to A_{600} = 0.4-0.6, and were then grown with shaking (200 rpm) at 20 °C for a further 16 h. For expression of $^{15}N^{13}C$ labelled CshA_RD13, a culture (100 mL) of *E. coli* BL21 (DE3) cells harbouring CshA_RD13::pOPINF was grown overnight with shaking at 37 °C. Cells were harvested by centrifugation, washed in resuspension buffer, and used to inoculate 2 L of M9 minimal media (50 mM KH_2PO_4, 25 mM Na_2HPO_4 (pH 6.8), 10 mM NaCl, 1 mM $MgSO_4$, 0.3 mM $CaCl_2$, 1 mg ml^{-1} biotin, 1 mg ml^{-1} thiamin), supplemented with carbenicillin (50 μg ml^{-1}), trace elements (5 ml per L, 100x), 0.5 g/L $^{15}NH_4Cl$ and 2 g/L ^{13}C glucose. The cells were grown to A_{600} = 0.8-0.9. Protein expression was induced by the addition of IPTG (1 mM), and the cell cultures transferred to 25 °C, with shaking at 200 rpm, and grown for a further 16 h.

Protein purification

All recombinant proteins were purified using the same general strategy. Cells were harvested by centrifugation and lysed. Cell debris was removed by centrifugation and the remaining supernatant liquids applied to a HiTrap Ni^{2+} affinity column (GE Healthcare). Proteins were eluted with an imidazole gradient of 10-500 mM over 15 column volumes. Fractions (2 mL) found to contain the target protein of interest (as identified by SDS-PAGE analysis) were pooled and concentrated. Protein samples were subjected to further purification using SEC by passage through either a Superdex 16/60 S75 column (CshA_RD1, CshA_RD3, CshA_RD5, CshA_RD7 and CshA_RD13) or a Superdex 16/60 S200 column

(CshA_RD1-17), both GE Healthcare. For unlabelled proteins SEC was performed in 50 mM Tris-HCl, 150 mM NaCl, pH 7.5. For labelled proteins SEC purification was performed in 20 mM phosphate, 50 mM NaCl, pH 7.5. Protein-containing fractions were pooled, concentrated to 20 mg ml^{-1}, and stored at 4 °C.

Analytical ultracentrifugation

Sedimentation velocity analytical ultracentrifugation (SV-AUC) experiments were performed using a Beckman Optima XL-I. Sedimentation of the CshA_RD1-17 was monitored at 40000 rpm and 20°C using the UV-visible absorption system at a wavelength of 280 nm. The sample concentration was 6.22 μM in buffer (20 mM Tris-HCl, 150 mM NaCl, pH 7.5). The sedimentation profiles were fitted in SEDFIT using the continuous distribution c(s) Lamm equation model. The partial specific volume of CshA_RD1-17 (0.7279 cm3 g^{-1}) was calculated from the primary sequence using SEDFIT. The density and viscosity of the buffer were measured using an Anton-Paar rolling-ball viscometer (Lovis 2000 M/ME) and found to be 1.002921 g cm^3 and 1.0218 mPa.s respectively.

Small angle X-ray scattering

Small angle X-ray scattering (SAXS) data of CshA_RD1-17 were collected at the Diamond Light Source synchrotron (beamline B21) with a fixed camera length configuration (4.014 m) at 12.4 keV. Size-exclusion chromatography coupled SAXS (SEC-SAXS) using an Agilent HPLC system was utilised to collect the data. The sample was measured at a concentration of 25.8 μM in buffer [20 mM Tris-HCl, 150 mM NaCl, 5 mM KNO_3, 1% sucrose, pH 7.5]. 2D scattering profiles were reduced using in-house software. The data were scaled, merged and background-subtracted using the ScÅtter software package (34). GNOM and BAYESAPP were used to generate pair distance distribution plots from the scattering curves. Form factor fitting was carried out with SASVIEW using a flexible cylinder model. The model describes a chain that is defined by the contour length (L) and the Kuhn length, b. The Kuhn length is defined as twice the persistence length, over which the chain can be described as rigid, and values above that expected for a random coil can be ascribed to the range of possible torsional angles between residues and to folded structural elements within the polypeptide. The contour length is the linearly extended length of the particle without stretching the backbone. For completely disordered chains behaving as a random coil, b is between 18-20 Å. The theoretical contour length for a fully disordered protein is 3.84 Å per residue and is defined by the number of residues and the spacing between Cα positions. EOM was to analyse the experimental data using the ensemble optimisation. RANCH was used to generate a pool of 10000 independent conformational models based on the primary sequence and homology models of folded RD domains. GAJOE was used to select an ensemble of models whose combined theoretical scattering profiles best approximated the measured data using a genetic algorithm.

Proteolytic His-tag cleavage

Following nickel affinity and size exclusion purification of recombinant CshA_RD1, CshA_RD3, CshA_RD5, CshA_RD7 and CshA_RD13 CshA proteins, their hexa-histadine tags were cleaved off by 3C protease digestion (Pierce). This was carried out according to manufacturer's protocol (Pierce): 3C protease (1 mg ml^{-1}) was incubated with his-tagged CshA protein (5 mg ml^{-1}) overnight at 4 °C with agitation. The cleaved CshA proteins were separated from the un-cleaved material by passage through a HiTrap Ni^{2+} affinity column (GE Healthcare) equilibrated with buffer (20 mM potassium phosphate, 100 mM NaCl, pH 7.0). Cleaved protein was eluted with 5 column volumes of the same buffer. Uncleaved protein was then eluted with elution buffer (20 mM potassium phosphate, 100 mM NaCl, 1 M imidazole, pH 7.0). Cleaved protein was concentrated to 5-10 mg ml^{-1}.

Equilibrium unfolding studies

Equilibrium unfolding studies were performed by monitoring the change in intrinsic tyrosine fluorescence as a consequence of increasing urea concentration. All spectra were collected using a Horiba - Jobin YVON Fluorolog. Protein concentrations of 10 μM in buffer (20 mM potassium phosphate, 100 mM NaCl, pH 7.0), plus varying concentrations of urea, were mixed and samples left to equilibrate for 1 h at 20°C prior to analysis. All fluorescence experiments were performed at 23 °C. For each sample an emission

spectrum was measured over the range 290–320 nm using an excitation wavelength of 278 nm. For analysis, the fluorescence intensity at 306 nm was plotted as a function of urea concentration and data were fitted to a two-state equilibrium unfolding model.

NMR spectroscopy

NMR datasets were collected at 20°C, utilising a Varian VNMRS 600 MHz spectrometer with a cryogenic cold-probe. All NMR data were processed using NMRPipe (35). ^{1}H-^{15}N HSQC, HNCACB, CBCA(CO)NH, HNCA, HNHA, HN(CO)CA, HNCO, HN(CA)CO, C(CO)NH, HCCH-TOCSY, ^{15}N-TOCSY-HSQC, ^{15}N-NOESY-HSQC, ^{13}C-NOESY-HSQC and aromatic ^{13}C-NOESY-HSQC (150 ms mixing time) experiments were collected. A hydrogen-deuterium (HD) exchange experiment was conducted by recording ^{1}H-^{15}N HSQC experiments at several intervals following dissolution of freeze-dried protein in D_2O. 2D ^{1}H-^{1}H TOCSY and NOESY experiments were recorded on the fully exchanged protein sample. ^{15}N-NOESY-HSQC and ^{13}C-NOESY-HSQC spectra (150 ms mixing time) were also recorded at 20°C on a Varian INOVA 900 MHz spectrometer with a cryogenic cold-probe (Henry Wellcome Building for NMR, University of Birmingham). Proton chemical shifts were referenced with respect to the water signal relative to DSS. Spectra were assigned using CcpNmr Analysis 2.4 (36). Structure calculations were conducted using ARIA 2.3 (37). 20 structures were calculated at each iteration except iteration 8, where 200 structures were calculated. The 20 lowest energy structures from this iteration went on to be water refined and the 15 lowest energy structures chosen as a representative ensemble. Network anchoring was used during iterations 0, 1 and 2, and all iterations were corrected for spin diffusion (38). Two cooling phases, each with 8000 steps were used. Torsion angle restraints were calculated using both TALOS+ (39) and DANGLE (40). Restraints were included for residues where both programs gave an unambiguous result in the same area of the Ramachandran plot. The restraints were based on those provided by DANGLE but extended if the TALOS+ restraints went beyond these. This process resulted in slightly fewer, looser restraints than either program on their own, but aimed to reduce the number of over-restrained angles. χ1 angle restraints were introduced for Val2107 and Val2109 as the orientation of these side chains was clearly defined by their NOE pattern, though the selection of structures based on global energy scores meant that not all structures resulted in these orientations unless these restraints were introduced. The HD exchange experiment showed 28 NH groups to be protected after 8 minutes, including two Gln side-chain amides (see Figure S4). In addition, NOEs were observed to a Thr Hγ1 hydrogen, suggesting that this was also involved in a hydrogen bond. Initial structure calculations were conducted without hydrogen bond restraints. Hydrogen bond donors were then identified, and corresponding hydrogen bond restraints included in later calculations. Structures were validated using the Protein Structure Validation Software (PSVS) suite 1.5 (41) and CING (42).

Data Availability

All data presented in this paper are contained within the manuscript and supporting information. NMR restraints and chemical shifts for CshA_RD13 have been deposited with the PDB with accession code 6SZC.

Acknowledgements

This work was funded in part by NIH grant DE016690, BBSRC grants BB/L01386X/1 and BB/M025624/1, and through the award of a Royal Society University Research Fellowship to P.R.R. (UF080534). K.L.V. was supported by an EPSRC BCFN PhD studentship (EP/G036780/1). We thank Drs Sara Whittaker (University of Birmingham) and Roz Ellis (University of Bristol) for assistance with NMR data collection, Dr Paul Curnow (University of Bristol) for assistance with CD data analysis, and Dr Robert Rambo (Diamond Light Source) for assistance with SAXS data collection.

Conflict of Interest

The authors declare that they have no conflicts of interest with the contents of this article.

Author Contributions

C.R.B. and V.V.P. performed gene cloning and produced protein. V.A.H. and M.P.C. collected NMR data. C.R.B. and V.A.H. analyzed NMR data and performed structure calculations. C.R.B and K.L.V. collected and analyzed AUC and SAXS data. V.V.P., A.E.P and S.G.B. performed CD and unfolding studies. All authors analyzed data and wrote the manuscript. D.F., H.F.J., S.G.B., M.P.C., A.H.N., and P.R.R. conceived the study and directed the research.

REFERENCES

1. Green, J. L., Bohannan, B. J. M., and Whitaker, R. J. (2008) Microbial biogeography: From taxonomy to traits. *Science* **320**, 1039-1043
2. Martiny, J. B., Bohannan, B. J., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Ovreas, L., Reysenbach, A. L., Smith, V. H., and Staley, J. T. (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**, 102-112
3. Kline, K. A., Falker, S., Dahlberg, S., Normark, S., and Henriques-Normark, B. (2009) Bacterial adhesins in host-microbe interactions. *Cell Host Microbe* **5**, 580-592
4. Pizarro-Cerda, J., and Cossart, P. (2006) Bacterial adhesion and entry into host cells. *Cell* **124**, 715-727
5. Wizemann, T. M., Adamou, J. E., and Langermann, S. (1999) Adhesins as targets for vaccine development. *Emerg Infect Dis* **5**, 395-403
6. Klemm, P., Vejborg, R. M., and Hancock, V. (2010) Prevention of bacterial adhesion. *Appl Microbiol Biotechnol* **88**, 451-459
7. Proft, T., and Baker, E. N. (2009) Pili in Gram-negative and Gram-positive bacteria - structure, assembly and their role in disease. *Cell Mol Life Sci* **66**, 613-635
8. Allen, W. J., Phan, G., and Waksman, G. (2012) Pilus biogenesis at the outer membrane of Gram-negative bacterial pathogens. *Curr Opin Struct Biol* **22**, 500-506
9. Kang, H. J., Coulibaly, F., Clow, F., Proft, T., and Baker, E. N. (2007) Stabilizing isopeptide bonds revealed in gram-positive bacterial pilus structure. *Science (New York, N Y)* **318**, 1625-1628
10. Larson, M. R., Rajashankar, K. R., Patel, M. H., Robinette, R. A., Crowley, P. J., Michalek, S., Brady, L. J., and Deivanayagam, C. (2010) Elongated fibrillar structure of a streptococcal adhesin assembled by the high-affinity association of alpha- and PPII-helices. *Proc Natl Acad Sci U S A* **107**, 5983-5988
11. Macintosh, R. L., Brittan, J. L., Bhattacharya, R., Jenkinson, H. F., Derrick, J., Upton, M., and Handley, P. S. (2009) The terminal A domain of the fibrillar accumulation-associated protein (Aap) of Staphylococcus epidermidis mediates adhesion to human corneocytes. *J Bacteriol* **191**, 7007-7016
12. Rego, S., Heal, T. J., Pidwill, G. R., Till, M., Robson, A., Lamont, R. J., Sessions, R. B., Jenkinson, H. F., Race, P. R., and Nobbs, A. H. (2016) Structural and Functional Analysis of Cell Wall-anchored Polypeptide Adhesin BspA in Streptococcus agalactiae. *J Biol Chem* **291**, 15985-16000
13. Jameson, M. W., Jenkinson, H. F., Parnell, K., and Handley, P. S. (1995) Polypeptides associated with tufts of cell-surface fibrils in an oral Streptococcus. *Microbiology* **141**, 2729-2738
14. Wu, H., Mintz, K. P., Ladha, M., and Fives-Taylor, P. M. (1998) Isolation and characterization of Fap1, a fimbriae-associated adhesin of *Streptococcus parasanguis* FW213. *Mol Microbiol* **28**, 487-500
15. Wu, H., and Fives-Taylor, P. M. (1999) Identification of dipeptide repeats and a cell wall sorting signal in the fimbriae-associated adhesin, Fap1, of Streptococcus parasanguis. *Mol Microbiol* **34**, 1070-1081

16. Froeliger, E. H., and Fives-Taylor, P. (2001) Streptococcus parasanguis fimbria-associated adhesin fap1 is required for biofilm formation. *Infect Immun* **69**, 2512-2519
17. McNab, R., Forbes, H., Handley, P. S., Loach, D. M., Tannock, G. W., and Jenkinson, H. F. (1999) Cell wall-anchored CshA polypeptide (259 kilodaltons) in Streptococcus gordonii forms surface fibrils that confer hydrophobic and adhesive properties. *J Bacteriol* **181**, 3087-3095
18. McNab, R., Holmes, A. R., Clarke, J. M., Tannock, G. W., and Jenkinson, H. F. (1996) Cell surface polypeptide CshA mediates binding of Streptococcus gordonii to other oral bacteria and to immobilized fibronectin. *Infect Immun* **64**, 4204-4210
19. Holmes, A. R., McNab, R., and Jenkinson, H. F. (1996) Candida albicans binding to the oral bacterium Streptococcus gordonii involves multiple adhesin-receptor interactions. *Infect Immun* **64**, 4680-4685
20. Jakubovics, N. S., Brittan, J. L., Dutton, L. C., and Jenkinson, H. F. (2009) Multiple adhesin proteins on the cell surface of Streptococcus gordonii are involved in adhesion to human fibronectin. *Microbiology* **155**, 3572-3580
21. Back, C. R., Sztukowska, M. N., Till, M., Lamont, R. J., Jenkinson, H. F., Nobbs, A. H., and Race, P. R. (2017) The Streptococcus gordonii Adhesin CshA Protein Binds Host Fibronectin via a Catch-Clamp Mechanism. *J Biol Chem* **292**, 1538-1549
22. Pakula, R., and Walczak, W. (1963) On the nature of competence of transformable streptococci. *Journal of general microbiology* **31**, 125-133
23. Berrow, N. S., Alderton, D., Sainsbury, S., Nettleship, J., Assenberg, R., Rahman, N., Stuart, D. I., and Owens, R. J. (2007) A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic Acids Res* **35**, e45
24. Borgia, M. B., Borgia, A., Best, R. B., Steward, A., Nettels, D., Wunderlich, B., Schuler, B., and Clarke, J. (2011) Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* **474**, 662-665
25. Smith, B. O., Picken, N. C., Westrop, G. D., Bromek, K., Mottram, J. C., and Coombs, G. H. (2006) The structure of Leishmania mexicana ICP provides evidence for convergent evolution of cysteine peptidase inhibitors. *The Journal of biological chemistry* **281**, 5821-5828
26. Gruszka, D. T., Wojdyla, J. A., Bingham, R. J., Turkenburg, J. P., Manfield, I. W., Steward, A., Leech, A. P., Geoghegan, J. A., Foster, T. J., Clarke, J., and Potts, J. R. (2012) Staphylococcal biofilm-forming protein has a contiguous rod-like structure. *Proc Natl Acad Sci U S A* **109**, E1011-1018
27. Gruszka, D. T., Whelan, F., Farrance, O. E., Fung, H. K. H., Paci, E., Jeffries, C. M., Svergun, D. I., Baldock, C., Baumann, C. G., Brockwell, D. J., Potts, J. R., and Clarke, J. (2015) Cooperative folding of intrinsically disordered domains drives assembly of a strong elongated protein. *Nat Commun* **6**, 7271
28. Gruszka, D. T., Mendonca, C. A. T. F., Paci, E., Whelan, F., Hawkhead, J., Potts, J. R., and Clarke, J. (2016) Disorder drives cooperative folding in a multidomain protein. *Proc Natl Acad Sci U S A* **113**, 11841-11846
29. Formosa-Dague, C., Speziale, P., Foster, T. J., Geoghegan, J. A., and Dufrene, Y. F. (2016) Zinc-dependent mechanical properties of Staphylococcus aureus biofilm-forming surface protein SasG. *Proc Natl Acad Sci U S A* **113**, 410-415
30. Troffer-Charlier, N., Ogier, J., Moras, D., and Cavarelli, J. (2002) Crystal structure of the V-region of Streptococcus mutans antigen I/II at 2.4 A resolution suggests a sugar preformed binding site. *J Mol Biol* **318**, 179-188
31. Forsgren, N., Lamont, R. J., and Persson, K. (2009) Crystal structure of the variable domain of the Streptococcus gordonii surface protein SspB. *Protein Sci* **18**, 1896-1905
32. Forsgren, N., Lamont, R. J., and Persson, K. (2010) Two intramolecular isopeptide bonds are identified in the crystal structure of the Streptococcus gordonii SspB C-terminal domain. *J Mol Biol* **397**, 740-751

33. Larson, M. R., Rajashankar, K. R., Crowley, P. J., Kelly, C., Mitchell, T. J., Brady, L. J., and Deivanayagam, C. (2011) Crystal structure of the C-terminal region of Streptococcus mutans antigen I/II and characterization of salivary agglutinin adherence domains. *The Journal of biological chemistry* **286**, 21657-21666
34. Forster, S., Apostol, L., and Bras, W. (2010) Scatter: software for the analysis of nano- and mesoscale small-angle scattering. *J Appl Crystallogr* **43**, 639-646
35. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes. *J Biomol Nmr* **6**, 277-293
36. Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, P., Ulrich, E. L., Markley, J. L., Ionides, J., and Laue, E. D. (2005) The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins* **59**, 687-696
37. Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T. E., and Nilges, M. (2007) ARIA2: Automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics* **23**, 381-382
38. Linge, J. P., Habeck, M., Rieping, W., and Nilges, M. (2004) Correction of spin diffusion during iterative automated NOE assignment. *J Magn Reson* **167**, 334-342
39. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009) TALOS plus : a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol Nmr* **44**, 213-223
40. Cheung, M. S., Maguire, M. L., Stevens, T. J., and Broadhurst, R. W. (2010) DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. *J Magn Reson* **202**, 223-233
41. Bhattacharya, A., Tejero, R., and Montelione, G. T. (2007) Evaluating protein structures determined by structural genomics consortia. *Proteins* **66**, 778-795
42. Doreleijers, J. F., da Silva, A. W. S., Krieger, E., Nabuurs, S. B., Spronk, C. A. E. M., Stevens, T. J., Vranken, W. F., Vriend, G., and Vuister, G. W. (2012) CING: an integrated residue-based structure validation program suite. *J Biomol Nmr* **54**, 267-283
43. Luthy, R., Bowie, J. U., and Eisenberg, D. (1992) Assessment of Protein Models with 3-Dimensional Profiles. *Nature* **356**, 83-85
44. Laskowski, R. A., Macarthur, M. W., Moss, D. S., and Thornton, J. M. (1993) Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *J Appl Crystallogr* **26**, 283-291
45. Lovell, S. C., Davis, I. W., Adrendall, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003) Structure validation by C alpha geometry: phi,psi and C beta deviation. *Proteins-Structure Function and Genetics* **50**, 437-450

FOOTNOTES

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

The abbreviations used are: CshA, cell surface hydrophobicity protein A; RD_13, repeat domain 13; SEC, size exclusion chromatography; SAXS, small angle X-ray scattering; SV-AUC, sedimentation velocity analytical ultracentrifugation;

TABLES

Domain	$\Delta G_{D-N}^{H_2O}$ $(kcal\ mol^{-1})$	m_{D-N} $(kcal\ mol^{-1}\ M^{-1})$
RD1	-0.64	0.66
RD3	-1.53	1.17
RD5	-	-
RD7	-1.47	1.24
RD13	-3.42	1.74

Table 1. Folding data for CshA repeat domains.

Degree of assignment[a]	
Backbone (Cα, C′, N and H^N) (%)	85.0
Side-chain H (%)	88.4
Side-chain non-H (%)	74.5
Number of restraints	
Distance restraints	
Intra-residue (|i-j| = 0)	489
Sequential (|i-j| = 1)	307
Medium range (2 ≤ |i-j| < 5)	165
Long range (|i-j| ≥ 5)	369
Ambiguous	470
Total	1800
Hydrogen bond restraints	29
Dihedral angle restraints (Φ/Ψ/χ^1)	120 (59/59/2)
Restraint statistics[b]	
r.m.s. of distance violations (Å)	0.018 ± 0.002
r.m.s. of dihedral violations (°)	0.57 ± 0.09
violations > 0.5 Å	0
violations > 0.3 Å	1.3 ± 1.5
violations > 0.1 Å	14.2 ± 2.9
r.m.s. from idealised covalent geometry[b]	
Bonds (Å)	0.0041 ± 0.0002
Angles (°)	0.53 ± 0.03
Impropers (°)	1.87 ± 0.14
Structural quality	
Ramachandran statistics[c]	
Most favoured regions (%)	80.1[a] / 84.6[d]
Allowed regions (%)	18.9[a] / 15.4[d]
Generously allowed regions (%)	0.3[a] / 0.0[d]
Disallowed regions (%)	0.7[a] / 0.0[d]
CING % ROG scores (R/O/G)[a] [42]	27 / 18 / 55
Verify3D Z-score [43]	-2.41[a]
Prosa II Z-score [41]	-0.54[a]
Procheck Z-score (Φ/Ψ) [44]	-3.38[a] / -2.75[d]
Procheck Z-score (all) [44]	-4.97[a] / -4.55[d]
MolProbity Z-score [45]	-1.13[a]
No. of close contacts	11[a,e]
Coordinates precision (rmsd)	
All backbone atoms (Å)	1.2[a] / 0.6[d] / 0.4[f]
All heavy atoms (Å)	1.6[a] / 1.0[d] / 0.7[f]

[a] residues 2053-2130
[b] values reported by ARIA 2.3 [37]
[c] values reported by Procheck [44]
[d] ordered residues (2056-2099, 2104-2124) as calculated by PSVS 1.5 [41]
[e] value reported by PDB validation software
[f] residues in secondary structure as calculated by PSVS 1.5 [41]
(2067-2069, 2075-2078, 2082-2087, 2091-2096, 2107-2112, 2118-2123)

Table 2. NMR Assignment, Structure Calculation and Validation statistics for CshA_RD13.

Parameter		
Repeat domains in model	1-17	1, 3-4, 7-8, 14-16
Ensemble R_g	99.04	120.0
Ensemble D_{max}	300.94	353.34
Fit quality (χ^2)	2.720	3.008

Table 3. EOM fitting parameters for different models.

FIGURES

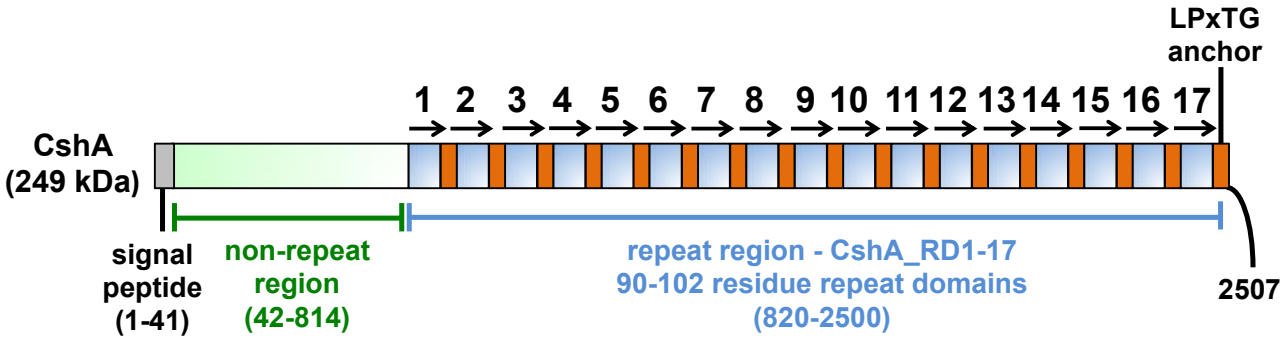

Figure 1. Schematic representation of CshA. Individual repeat domains (RDs) are highlighted in blue with proposed interdomain linkers highlighted in red. The adhesive non-repeat region is highlighted in green.

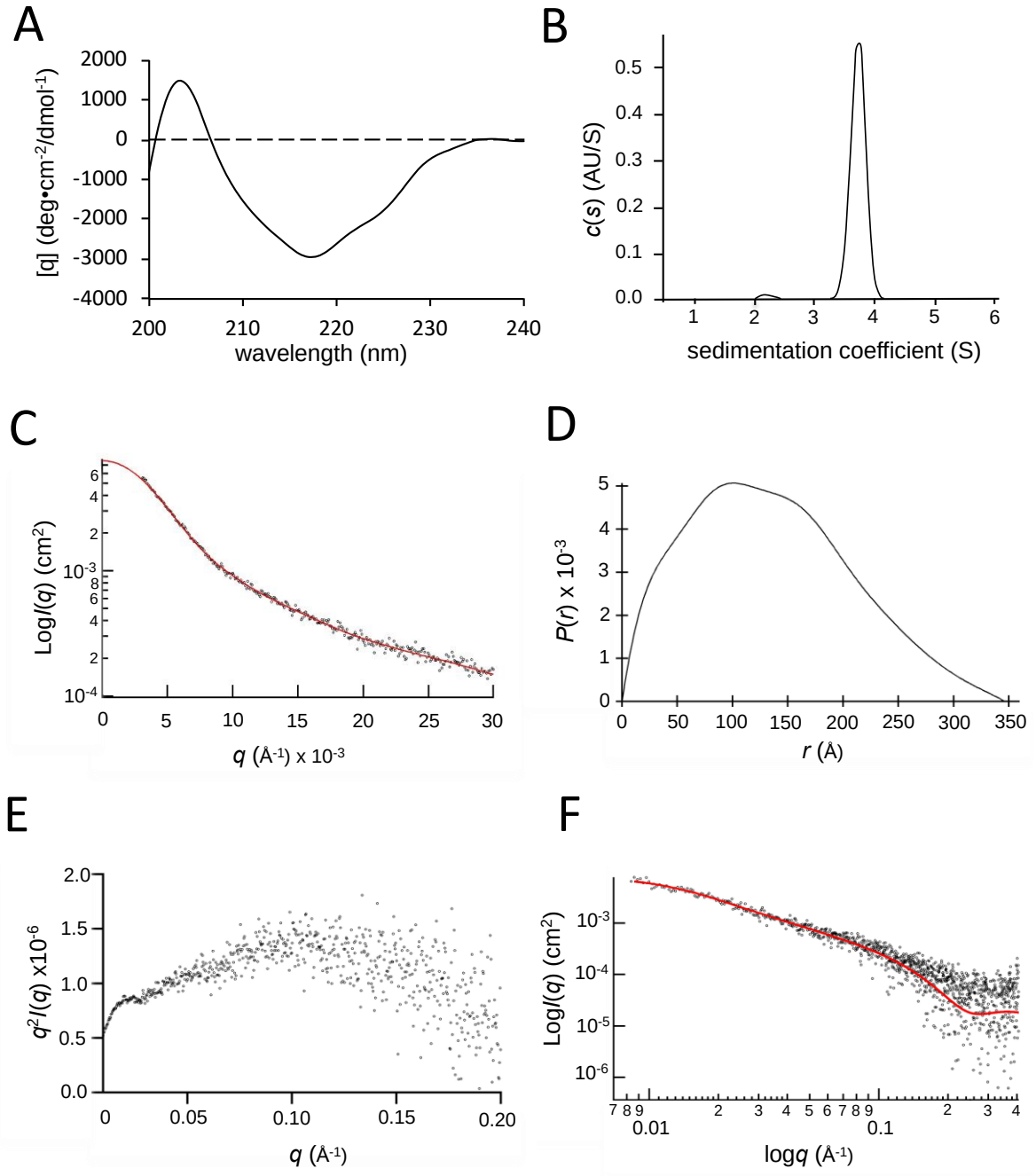

Figure 2. The intact CshA repeat region adopts an extended polymeric structure in solution: (*A*) Far-UV CD spectrum of CshA_RD1-17. (*B*) Sedimentation velocity AUC of CshA_RD1-17. (*C*) SAXS profile for CshA_RD1-17 (black circles) and inverse Fourier transform fit for *P*(*r*) distribution (red line). (D) *P*(*r*) distribution for CshA_RD1-17 derived from the scattering profile shown in (*C*). (E) Kratky plot derived from the scattering profile shown in (*C*). (*F*) SAXS data for CshA_RD1-17 fitted to a flexible cylinder model.

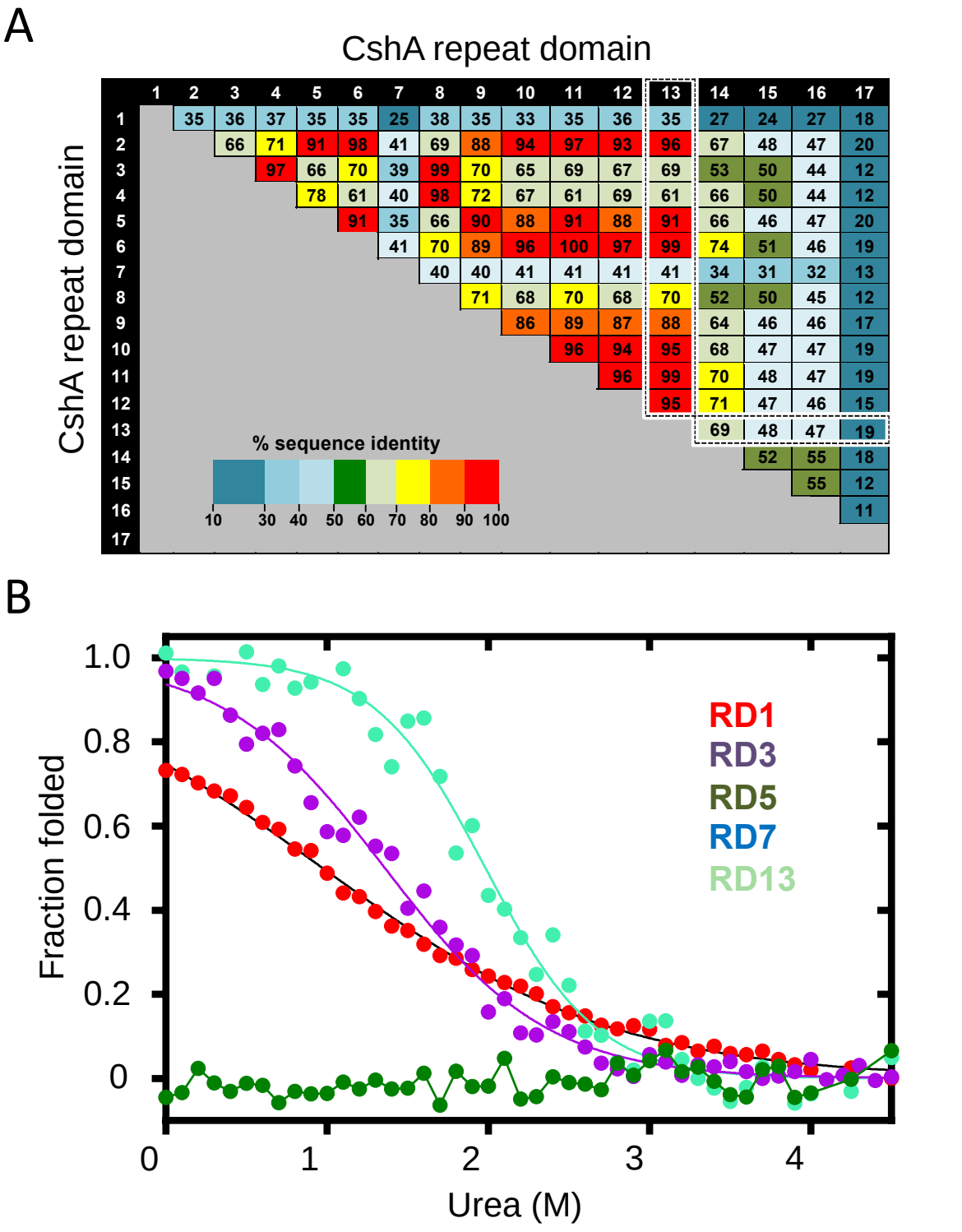

Figure 3. Individual CshA repeat domains exhibit varying stabilities and unfolding behaviours: (*A*) Heat plot showing primary amino acid sequence identities between pairs of CshA repeat domains. Values correspond to % sequence identities. Numbered axes refer to individual CshA repeat domains. (*B*) Urea-induced equilibrium unfolding transitions for individual CshA repeat domains.

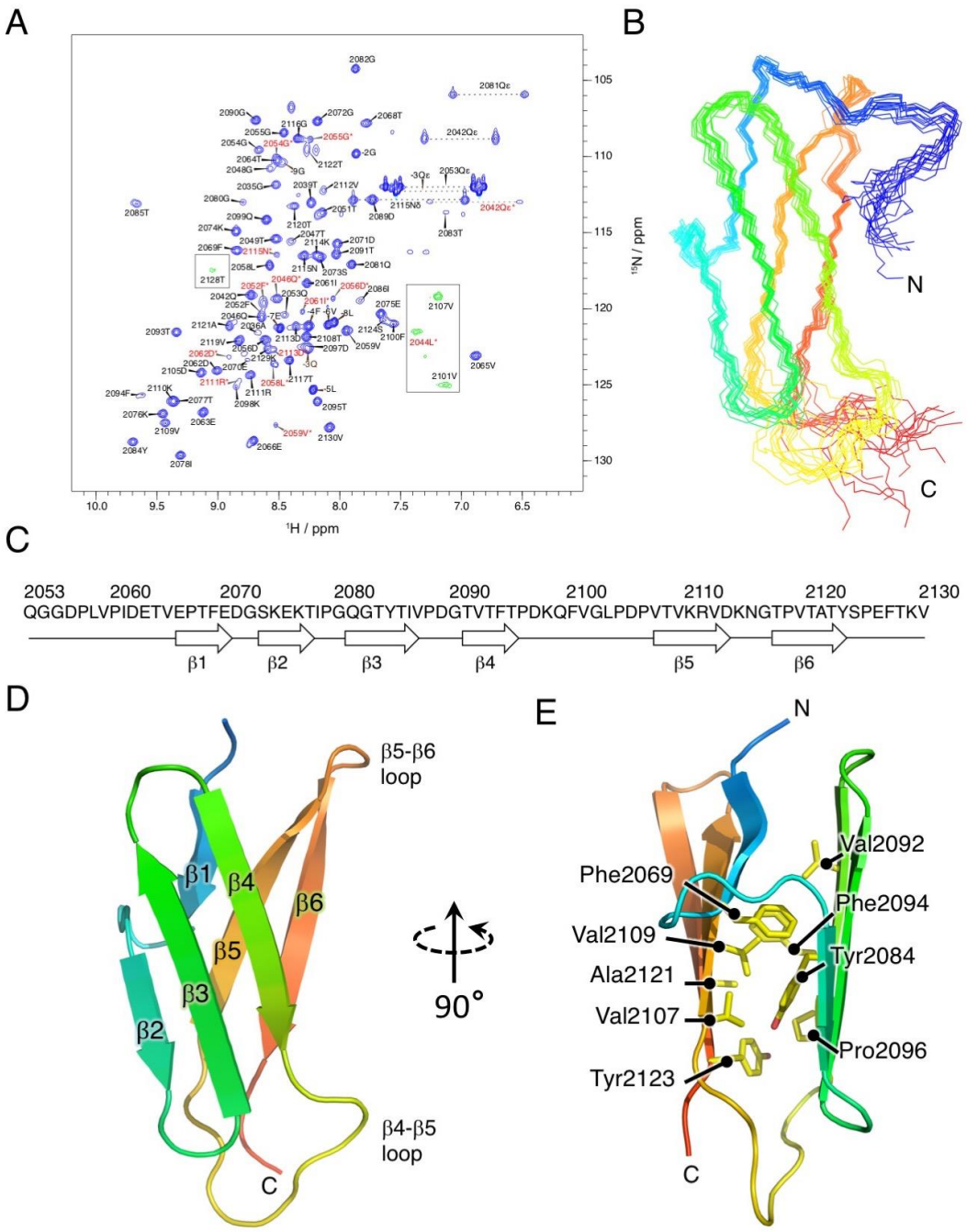

Figure 4. Solution structure of CshA_ RD13 reveals a new protein fold: (*A*) ^{1}H-^{15}N HSQC spectrum of CshA_RD13 recorded at 20°C and 14.1T (600 MHz). Boxed contours in green are shown at a lower contour level to the rest of the spectrum. Starred peak labels in red indicate additional minor conformations, not included in the structure calculations. Peak labels with negative peak numbers indicate peaks from the unstructured His-tag. (*B*) Ensemble of 15 lowest energy structures of CshA_RD13. (*D*) Sequence of CshA_RD13 depicted in *B*, *D* and *E*. (*D*) Cartoon diagram of CshA_RD13 with secondary structure elements labelled. (*E*) Structure of CshA_RD13 highlighting the composition of the hydrophobic core of the protein.

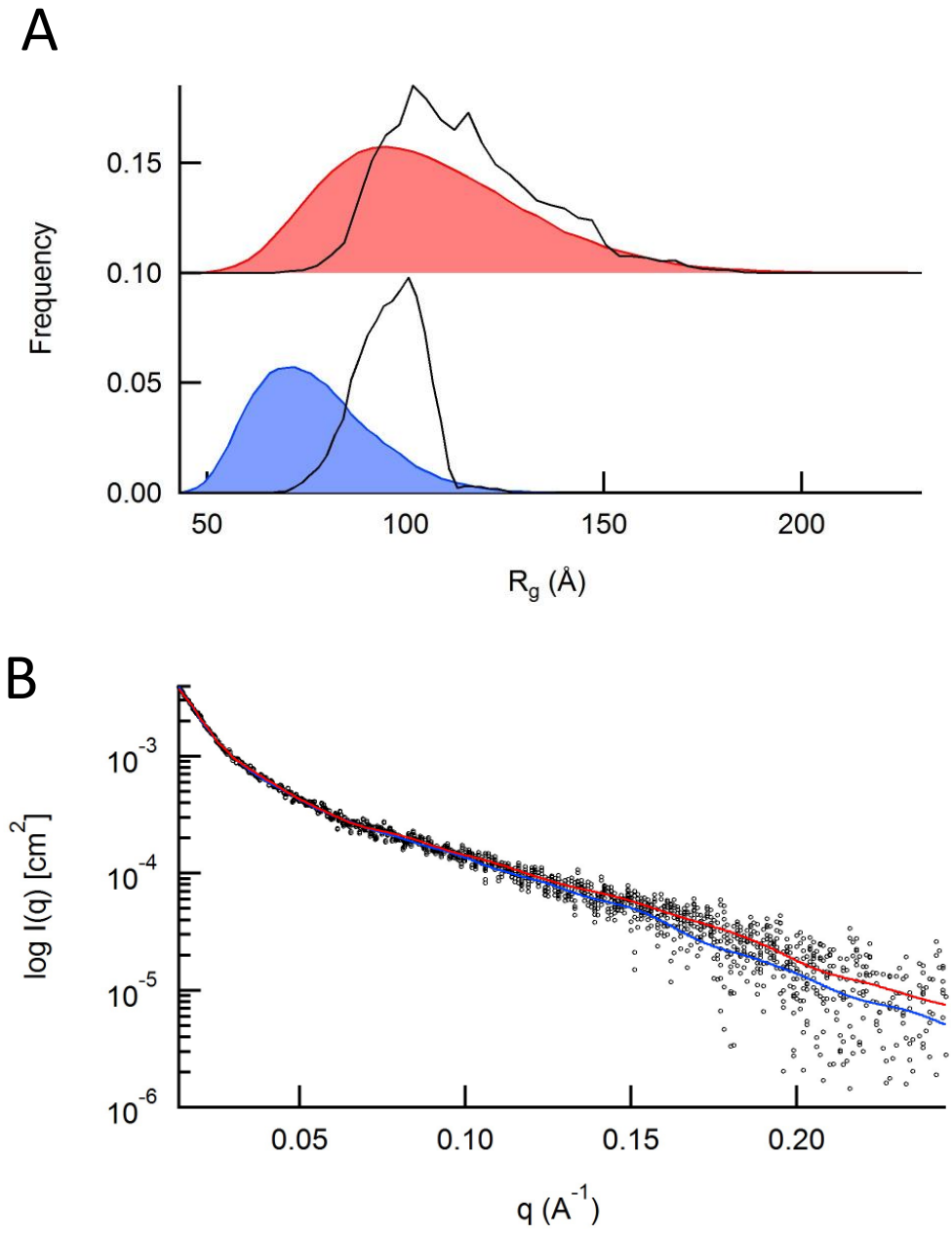

Figure 5. CshA_RD1-17 adopts a transitory dynamic structure comprising alternating regions of order and disorder: (*A*) Rg distributions for CshA_RD1-17 modelled with 17 globular repeat domains (blue) and 8 globular repeat domains (red). The Rg distribution of the full pool of structures generated with RANCH is shown in solid colour and the Rg distribution of the selected ensemble is shown in black. (B) EOM fits for CshA_RD1-17 modelled with 17 globular repeat domains (blue) and 8 globular repeat domains (red).