Title

Insight into the autoproteolysis mechanism of the RsgI9 anti-σ factor from Clostridium thermocellum.

Journal

Proteins: Structure, Function, and Bioinformatics, 92(8)

Authors

Takayesu, Allen
Mahoney, Brendan
Goring, Andrew
et al.

Publication Date

2024-08-01

DOI

10.1002/prot.26690

Peer reviewed

HHS Public Access

Author manuscript

Proteins. Author manuscript; available in PMC 2025 February 01.

Published in final edited form as:

Proteins. 2024 August ; 92(8): 946–958. doi:10.1002/prot.26690.

Insight into the autoproteolysis mechanism of the RsgI9 anti-σ factor from Clostridium thermocellum

Allen Takayesu[a,b], **Brendan J. Mahoney**[b,c], **Andrew K. Goring**[a,b], **Tobie Jessup**[a], **Rachel R Ogorzalek Loo**[a,c], **Joseph A. Loo**[a,b,c], **Robert T. Clubb**[a,b,c,d]

[a]Department of Chemistry and Biochemistry, Los Angeles, 611 Charles Young Drive East, Los Angeles, CA 90095, USA.

[b]UCLA-DOE Institute of Genomics and Proteomics, Los Angeles, 611 Charles Young Drive East, Los Angeles, CA 90095, USA.

[c]Molecular Biology Institute. University of California, Los Angeles, 611 Charles Young Drive East, Los Angeles, CA 90095, USA.

Abstract

Clostridium thermocellum is a potential microbial platform to convert abundant plant biomass to biofuels and other renewable chemicals. It efficiently degrades lignocellulosic biomass using a surface displayed cellulosome, a megadalton sized multienzyme containing complex. The enzymatic composition and architecture of the cellulosome is controlled by several transmembrane biomass-sensing RsgI-type anti-σ factors. Recent studies suggest that these factors transduce signals from the cell surface via a Conserved RsgI Extracellular (CRE) domain (also called a periplasmic domain) that undergoes autoproteolysis through an incompletely understood mechanism. Here we report the structure of the autoproteolyzed CRE domain from the *C. thermocellum* RsgI9 anti-σ factor, revealing that the cleaved fragments forming this domain associate to form a stable α/β/α sandwich fold. Based on AlphaFold2 modeling, molecular dynamics simulations and tandem mass spectrometry, we propose that a conserved Asn-Pro bond in RsgI9 autoproteolyzes via a succinimide intermediate whose formation is promoted by a conserved hydrogen bond network holding the scissile peptide bond in a strained conformation. As other RsgI anti-σ factors share sequence homology to RsgI9, they likely autoproteolyze through a similar mechanism.

Keywords

RsgI; Cellulosome; Anti-sigma factor; autoproteolysis; *Clostridium thermocellum* ; Conserved RsgI Extracellular (CRE) domain; Periplasmic domain

[d]To whom correspondence should be addressed: Robert T. Clubb. Department of Chemistry and Biochemistry, University of California, Los Angeles, 611 Charles Young Drive East, Los Angeles, CA 90095, USA; Fax (+1) 310 206 4779; rclubb@mbi.ucla.edu.

CONFLICT OF INTEREST

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

Introduction

The effects of climate change and finite petroleum supplies have created a pressing need to cost-effectively produce biofuels, chemicals, and materials from renewable and carbon-neutral sources [1]. Lignocellulosic plant biomass (LCB) is a promising feedstock to generate these biomaterials as it is the largest source of carbon in the biosphere (~450 gigatons) and can be produced sustainably [2]. Currently, the use of LCB as a feedstock is limited by its recalcitrance to degradation [3]. One approach to overcome this problem is to employ *Clostridium thermocellum* (also known as *Acetivibrio thermocellus* or *Hungateiclostridium thermocellum*) to break down biomass as this thermophilic obligate anaerobe possesses extremely potent cellulolytic activity [4, 5]. *C. thermocellum* has evolved the capacity to degrade LCB by displaying cellulosomes - massive multienzyme complexes that efficiently hydrolyze cellulose into its component sugars [6–10]. The high cellulolytic activity of the cellulosome originates from its modular architecture, which is capable of colocalizing a large a number of glycoside hydrolases (GHs) that work together to degrade cellulose and hemicellulose fibers within LCB [11]. In addition to *C. thermocellum,* a number of mesophilic and thermophilic microbial species have been shown to produce cellulosomes that confer potent cellulolytic activity [12, 13].

To degrade different types of biomass, *C. thermocellum* regulates the transcription of ~50 genes encoding GH enzymes that are incorporated into the cellulosome [14, 15]. Their expression is regulated by extracytoplasmic function (ECF) σ-factors, which together with their cognate anti-σ factors are responsible for altering GH gene expression in response to the presence of different types of extracellular polysaccharides [16]. A total of nine ECF σ-factors may regulate gene expression in response to different types of polysaccharides. These include eight SigI-type σ-factors (σ_1 to σ_8) whose activities are modulated by at least eight RsgI-type anti-σ factors (RsgI1–8) that are co-transcribed from the same operon, and σ_{24} which is regulated by the Rsi24c anti-σ factor [16–19]. In the signaling process for the SigI-type σ-factors, each of the membrane-bound anti-σ factors are believed to bind to different types of extracellular polysaccharides [17, 20–22]. They then transduce this signal across the membrane, triggering the release of their cognate SigI-type σ-factor from the cytoplasmic membrane [23]. The SigI-type σ-factor then binds to the RNA polymerase complex enabling it to transcribe specific GH encoding genes, as well as genes that encode scaffoldin proteins used to construct the cellulosome [24]. Each of the RsgI anti-σ factors contains a conserved cytoplasmic N-terminal domain (NTD) that is located in the cytoplasm, where it binds to its cognate σ-factor. This is followed by a single membrane embedded helix and an extracellularly located Conserved RsgI Extracellular domain (CRE domain). The CRE domain has also been referred to as either the periplasmic juxtaposition domain or SEAL domain [25, 26]. Following this segment is a non-conserved C-terminal region that frequently mediates polysaccharide binding. Biochemical studies have deduced the binding activities of several of the RsgI anti-σ factors. For example, RsgI1, RsgI2 and RsgI4 contain family 3 type carbohydrate binding modules (CBM3) known to interact with cellulose [17], RsgI3 recognizes pectin via protective antigen 14 domains, RsgI5 contains a family-10 CBM that binds to arabinose, and RsgI6 contains a family-10 GH module that degrades cellulose and xylan [17, 20, 21]. Recent studies of the RsgI-like factors from *Bacillus subtilis* and

C. thermocellum suggest that signaling occurs via a regulated intramembrane proteolysis (RIP) mechanism in which the anti-σ factor undergoes proteolytic cleavage at two sites, called site-1 and site-2 [25–27]. Site-1 cleavage occurs within their CRE domains as a result of autoproteolysis, predisposing the anti-σ factor for signal-induced structural changes that expose site-2 for degradation by the RasP protease in *B. subtilis* or RseP in *C. thermocellum*. The second cleavage event releases the cognate σ-factor from the RsgI's cytoplasmic NTD enabling it to selectively transcribe specific genes by targeting the RNA polymerase complex to their promoters.

Recently we reported the structure of the cellulose-binding ectodomain from the *C. thermocellum* RsgI9. It is an 'orphan' anti-σ factor [22] because unlike other RsgI factors, the gene encoding RsgI9 is not located in an operon that also contains a gene encoding a σ-factor. The ectodomain in RsgI9 adopts an elongated conformation in which a C-terminal bi-domain unit is likely projected from the cell surface to engage cellulose, while its conserved CRE domain resides near the membrane to presumably affect signal transduction. Using a combination of NMR spectroscopy, AlphaFold2, molecular dynamics (MD) simulations, and mass spectrometry we show that RsgI9 exists in a pre-cleaved state on the extracellular membrane that may enable it to transduce signals caused by carbohydrate binding into gene expression changes that alter the composition of the cellulosome, and we provide evidence that is compatible with CRE domain autoproteolyzing via a conserved mechanism involving a succinimide intermediate.

Results and Discussion

RsgI9's CRE domain undergoes autoproteolysis.

CRE domains are conserved in a range of membrane embedded anti-σ factors and have recently been implicated in RIP-mediated signal transduction [23]. To gain insight into its function, we recombinantly produced a polypeptide encoding RsgI9's CRE domain (CRE, residues G167-K343 of RsgI9) (Fig. 1A). The final step of the CRE purification process involved the application of size exclusion chromatography (SEC). Interestingly, we noticed that even though the purified protein eluted from the column at a position consistent with its predicted molecular weight (~20 kDa, 177 amino acids) (Fig. 1B), based on SDS-PAGE analysis the material within this peak contains two polypeptide chains with masses of ~2.4 and ~17.6 kDa as confirmed by MALDI-TOF (Fig. 1C). To confirm the identity of the components, the two fragments were separated, digested with trypsin, subjected to reversed-phase HPLC and then analyzed via electrospray ionization tandem mass spectrometry (ESI-MS/MS) (Fig. 1D). This procedure revealed that the fragments corresponded to polypeptides containing residues G167-N188 and P189-K343 in CRE, which was consistent with the full CRE protein undergoing proteolysis at its N188-P189 peptide bond. Close analysis of SDS-PAGE separated proteins in *E. coli* cells overexpressing CRE did not reveal the presence of the intact polypeptide, suggesting that autoproteolysis occurs rapidly after the protein is translated.

NMR structure of the autoproteolyzed CRE domain.

Co-elution of the 2.4 and 17.6 kDa CRE cleavage fragments in SEC experiments suggests that they associate with one another. To investigate the nature of this interaction, a ^{13}C- and ^{15}N-labeled sample of CRE was produced and its chemical shifts were assigned using triple resonance NMR methods (Fig. 2A) [28]. CRE adopts a folded structure as evidenced by its well dispersed ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectrum. Notably, both fragments within the cleaved CRE domain are structurally ordered. This is evident from the heteronuclear $^{15}N\{^{1}H\}$ nuclear Overhauser (hetNOE) data, since residues spanning both fragments within the protein exhibit values >0.6 (Fig. 2B). It is further substantiated by a plot of the random coil index (RCI) values predicted by TALOS-N, which reveals that residues in both the short and long fragments are structured based on their backbone chemical shifts (residues G177-I341) (Fig. 2C) [29].

The solution structure of the autoproteolyzed CRE domain was determined using multidimensional heteronuclear NMR data and simulated annealing approaches [30, 31]. The structure is represented by an ensemble of 20 conformers that are compatible with 2,044 structural restraints (1,668 NOE distances, 301 dihedral angles, 74 hydrogen bonds) (Table 1).

Consistent with the hetNOE data (Fig. 2B), residues Y180-I337 are structurally ordered in the ensemble; their backbone atoms can be superimposed to the mean structure with a root mean square deviation (RMSD) of 0.57 Å (Fig. 3A). CRE adopts an α/β/α sandwich that is constructed from three subdomains: helices α1–3, strands β1–5 that form a central sheet, and helices α4–7 (Fig. 3C). The β-strands are arranged in an anti-parallel manner, with the exception of strands β4 and β5 that are parallel to one another and located at the edge of the sheet. The shorter CRE fragment (residues G167-N188) comprises strand β1 that pairs in an anti-parallel manner with strands β2 and β4 located in the longer fragment. The cleavage site residues lie within the cleaved hairpin loop between the β1 and β2 strands and are partially excluded from the solvent by the side chains of residues located within α1 (D210) and α6 (K336), as well as the side chain of Y241 located in the loop region connecting α1 and β4 (Fig. 3B, yellow).

The structure is similar to recently reported structures of CRE domains from the *B. subtilis* RsgI anti-σ factor, as well as the RsgI1, RsgI2 and RsgI6 proteins from *C. thermocellum* (Fig. 4A) [27]. These proteins share 32–41% sequence identity with RsgI9 and also undergo autoproteolysis of an Asn-Pro peptide bond (Fig. 4B). The structure of RsgI9 is most similar to RsgI2, as their backbone atoms can be superimposed with an RMSD of 1.95 Å. A comparison of RsgI9 to the different CRE domain structures reveals the presence of many conserved residues that are positioned near the cleaved peptide bond (Fig. 4B) [26, 27]. These include amino acids within the conserved D[I/V]NPS sequence that harbor the cleaved Asn-Pro peptide bond and several residues from the longer fragment that interact with these amino acids. For example, semi-conserved polar contacts originating from the side chains N208 and D210 to the cleaved proline residue are observed, as well as contacts from L250, S310, and G312 to residues that surround the broken peptide bond (Fig. 3D, 4B). This conservation suggests that all CRE domains may autoproteolyze, consistent with experimental studies of RsgI from *B. subtilis* (^{Bsub}RsgI) and select RsgI proteins from

C. thermocellum (^{Cthe}RsgI1–4, 6). A notable exception is the RsgI8 CRE domain, which replaces the Asn-Pro sequence with Asp-Ala and may not undergo autoproteolysis [26, 27].

A conserved hydrogen-bond network may impart strain onto the Asn-Pro peptide bond to promote its cleavage.

To gain insight into the mechanism of cleavage, we used AlphaFold2 to model the structure of an intact CRE domain that exists prior to autoproteolysis (Fig. 5A) [34]. The top 5 models describing the structure were accurately predicted based on their pLDDT scores of 91.8 – 93.4. The intact (AlphaFold2) and cleaved (NMR) structures are closely related, as their backbone atoms can be superimposed with an RMSD of 1.94 Å. In general, the structural differences originate from alterations in the conformations of protein loops and the presence of an extra helical turn located at the end of the α7 helix in the intact protein. However, several features in the intact protein suggest that the region containing the Asn-Pro peptide bond adopts a stressed conformation that may promote its cleavage. For an example, an analysis using the program MolProbity reveals that N188-P189 ω dihedral angle describing the geometry of the scissile peptide bond is 83.2°, instead of the canonical value of ±180° for an unstrained, trans-peptide bond [35]. The φ and ψ torsion angles for P189 in this region also reside in the disfavored region of the Ramachandran plot (−107.3, 174.1). Finally, a CaBLAM analysis reveals several geometric outliers involving C^{α} atoms for residues V187-P189 suggesting that they adopt an energetically unfavorable conformation (Fig. 5B) [36]. In marked contrast, the NMR structure of the cleaved polypeptide does not exhibit this conformational stress; presumably because of the additional conformational freedom that is afforded by rupture of the N188 and P189 bond.

Molecular dynamics (MD) simulations were performed to investigate the origin of conformational strain in the CRE domain. Motions in the solvated model of the intact protein were simulated for 200 ns using the CHARMM36 all-atom force field [37]. The coordinates in the simulation stabilized after ~50 ns and thereafter exhibited only small root mean square fluctuations (RMSFs). Prominently, residues within and surrounding the β1-β2 hairpin containing the Asn-Pro peptide bond exhibited only relatively small conformational fluctuations (~0.75 Å RMSF values) and remained strained through the entirety of the simulation. In particular, the ω dihedral angle for the Asn-Pro peptide exhibited a non-ideal average value of 148.1° ± 8.4° (instead of ~±180°) with even larger transient excursions that move it as far as 55° away from planarity. In contrast, for all other residues in the protein their ω dihedral angles retain values around ±176° and thus their peptide bonds remain in a lower energy, planar configuration. Residues surrounding the scissile bond also remained in their original, unfavorable conformations. An inspection of the trajectory suggests that an extensive hydrogen bond network may hold the Asn-Pro peptide bond in a strained conformation (Fig. 6B). As part of this network, we also observed a long-lived water molecule that for the majority of the trajectory maintained hydrogen bonding interactions between residues D186, N188, and S190. Within the run time of the simulation, the water formed at least one hydrogen bond contact with the aforementioned residues for 94% of the simulation. In particular, the water molecule appears to form hydrogen bonds with both the main chain amide and carbonyl of N188, which had a site occupancy of 45% during the remaining 150 ns runtime after the system had stabilized Additionally, N208 and R313 form

hydrogen bond contacts with both the amide side chain and main chain carbonyl of N188. Although longer time-scale simulations are needed to fully assess the stability of this region, this network may hold the Asn-Pro peptide bond in its strained conformation. This would raise the energy of the uncleaved substrate, effectively reducing the activation energy barrier required for proteolysis. The functional importance of the network is consistent with studies of the ^{Cthe}RsgI2 CRE domain since mutation of the equivalent residues in this protein impair its autoproteolysis, specifically D186, S190, N208, and R313 [27].

Spontaneous cleavage may occur via a succinimide intermediate.

In principle, cleavage of the twisted scissile bond of CRE could occur via at least three distinct mechanisms: (i) direct hydrolysis of the strained peptide bond, (ii) an N→O / N→S acyl rearrangement mechanism that is caused by a nucleophilic action of a proximal residue such as a hydroxyl bearing serine residue [38, 39], or (iii) via a succinimide intermediate mechanism whereby the Asn side chain (N188) with the Asn-Pro scissile peptide bond undergoes cyclization and cleavage via the formation of a succinimide intermediate (Scheme I) [40].

Proteolysis via direct hydrolysis (i) or an acyl rearrangement (ii) seems unlikely. Direct hydrolysis seems unlikely for several reasons. First, in proteins that hydrolyze peptide bonds the water molecule is typically activated for nucleophilic attack on the scissile bond by a nearby metal or general base [41]. In the structure of RsgI9 and other CRE proteins, no metal is present near the labile peptide bond. There are numerous conserved residues near the labile bond that could potentially function to activate a water molecule for hydrolysis (e.g. S190, N208, and R313 in RsgI9). However, results obtained by the Feng and Rudner groups have shown that variants which alter these residues exhibit only mildly reduced autoproteolytic activity, suggesting that they are unlikely to function as a general base [25–27]. Second, the MD simulation of the uncleaved AlphaFold2 CRE model revealed the presence of a long-lived water molecule that stabilizes the hydrogen-bond network and could mediate hydrolysis, but is positioned too far away from scissile peptide bond making it a poor candidate for a nucleophile [42]. However, we cannot exclude the possibility that the 200 ns simulation may have been of insufficient duration to capture the presence of a long-lived water molecule that could mediate hydrolysis. Third, the cleavage product containing C-terminal Asn residue exhibits an isoform that is consistent with a mechanism involving a succinimide intermediate (see below). The N→O / N→S acyl rearrangement mechanism (ii) also seems unlikely to be the cause of CRE domain autoproteolysis. This is the most commonly observed mechanism of protein autoproteolysis and involves an acyl rearrangement mediated by a nearby side chain (e.g. Thr, Ser, Cys, Asn) that acts as a nucleophile that attacks the carbonyl group within the scissile peptide bond [43]. Proteins that utilize this mechanism include self-cleaving proteins such as the FoxR anti-σ factor [44], GPCR autoproteolysis inducing (GAIN) domains [45], and several viral polyproteins including those from SARS-CoV-2 [46]. This acyl rearrangement is unlikely for CRE as mutagenesis of highly conserved residues near the peptide bond that could serve as potential nucleophiles in the ^{Cthe}RsgI2 protein (N208, D210, and R313) does not significantly ablate its autoproteolysis [27]. Indeed, prior work on ^{Cthe}RsgI2 demonstrated that double mutants in which more than one potential nucleophile is altered do not completely disrupt

autoproteolysis, making this mechanism unlikely. Similarly, mutations of residues proximal to the autocleavage (Y182, D186, E192, V201, N208) site in the CRE of ^{Bsub}RsgI made by the Rudner group did not completely prevent autocleavage but did lead to a similar phenotype as the Δ*rsgI* mutant [25]. According to the solution structure and MD simulation, these residues are either involved within the hydrogen bond network surrounding the autocleavage site or within the core of the domain. Thus, mutations to these core residues would likely destabilize the domain such that the RsgI protein would be more susceptible to site-2 processing and activation.

We posit that the non-enzymatic cleavage of the Asn-Pro bonds in CRE domains occurs via a succinimide mechanism shown in Scheme 1. This autoproteolysis mechanism has been observed in the structurally unrelated αA-crystallin [40], aquaporin 0, FlhB and SpaS proteins [47–50]. Bond cleavage starts with the cyclization of the Asn side chain when its δ-nitrogen atom attacks its backbone carbonyl group to form a *gem*-hydroxylamine transition state that subsequently collapses into a C-terminal succinimide intermediate when the Asn-Pro bond breaks [51]. The C-terminal succinimide containing the Asn is then hydrolyzed to create a racemic mixture of cleavage product that contains either a C-terminal Asn or Asp-amide/*iso*Asn residue (Scheme 1). This reaction is typically rare in proteins as it competes with much faster Asn deamidation reaction that involves Asn cyclization [52]. However, Asn-Pro peptide bonds are prone to autoproteolysis since this competing deamidation reaction is disfavored because the proline backbone imine is a poor nucleophile [53]. If cleavage in RsgI9 occurs via a succinimide intermediate, the short peptide cleavage fragment should contain both Asn and *iso*Asn at its C-terminus (Scheme 1). LC-MS/MS was performed on a sample of purified CRE domain that had been digested with trypsin. This procedure identified 14 unique peptides corresponding to the larger fragment and 1 peptide corresponding to the shorter fragment (85.2% and 90.9% sequence coverage, respectively). Interestingly, the extracted ion chromatogram (XIC) revealed that all the tryptic peptides eluted with a standard unimodal elution profile with the notable exception of a shorter fragment containing the C-terminal Asn (or *iso*Asn) residue which exhibited a bimodal profile (Fig. 7A). The identity of the bimodal peptide was confirmed by LC-MS/MS as spanning residues G169-N188 (Fig. 7B). Notably, MS analysis of eluate from the bimodal chromatographic peak revealed that both lobes corresponded to peptides of *m/z* 1081.49. This behavior is consistent with the elution of a C-terminal Asn/*iso*Asn peptide pair. To further investigate this issue, we performed parallel reaction monitoring (PRM) on these tryptic digest products. Consistent with assignment as the autoproteolytic *iso*Asn and Asn products, both lobes of the LC peak exhibited nearly identical MS/MS spectra. This behavior is expected given that their backbone amides are similarly susceptible to collisional fragmentation (Fig. 7C). It is also notable that the single long-lived water molecule, observed to stabilize the hydrogen bond network in MD simulations, is positioned to facilitate cleaving the succinimide intermediate. Poised to hydrogen bond to the main chain carbonyl of N188 (heavy atom separation 3.0 Å), the water molecule can donate a proton as the peptide bond is broken.

Conclusion

In this work, we demonstrate that the RsgI9 CRE domain undergoes autoproteolysis at a conserved Asn-Pro sequence to form a stably folded structure in which the proteolytic fragments remain associated. Coupled with prior findings that show RsgI9's C-terminal ectodomain binds cellulose, these results are consistent with it regulating the composition of the cellulosome via a RIP signal transduction mechanism [22]. In this process, polysaccharide binding to the RsgI9 domain would cause the pre-cleaved CRE domain to separate, thereby exposing it to degradation by the RasP intramembrane metalloprotease to release a yet to be identified σ-factor that controls gene transcription. This work builds upon previous studies of the CRE domain by providing insight into the potential mechanism for autoproteolysis. Additionally, we have confirmed that the CRE of the orphan RsgI9 undergoes autoproteolysis, implicating its involvement in extracellular sensing and transcription regulation despite its unknown cognate σ-factor. Future studies will be focused on identifying this σ-factor and testing whether polysaccharide binding triggers CRE domain dissociation. Our NMR, MD and MS/MS analyses in combination with recently reported mutagenesis data provide insight into how a conserved Asn-Pro peptide bond in CRE domains is rapidly autoproteolyzed. We propose that a conserved hydrogen bond network and bound water molecule promote cleavage by maintaining the scissile peptide bond in a strained conformation that reduces the energy between the pre-cleaved substrate and transition state associated with bond breakage. This is similar to the hydrogen bond-maintained conformational strain seen in SEA domains, which have recently been proposed Brogan et al. to function similarly to CRE [26]. There are over 1000+ CRE homologs that contain the conserved D[I/V]NPS cleavage sequence across many species of gram-positive bacteria, suggesting that this mechanism is widely employed by RsgI-type anti-σ factors to mediate signal transduction.

Materials and Methods

Cloning and expression of CRE.

The nucleotide sequence for *C. thermocellum* DSM 1313 CRE (residues G167-K343) was cloned into a pET-29b expression plasmid which contains an N-terminal 6xHis affinity tag and followed by a Tobacco Etch Virus (TEV) protease cleavage site. The plasmid was transformed into *E. coli* BL21 (DE3) cells and cultured at 37 °C in LB media containing 50 μg/mL kanamycin until an OD_{600} of 0.6 – 0.8 was reached. Cells were then induced with 1mM isopropyl β-D-thiogalactoside (IPTG) and incubated with agitation at 17 °C for 12–17 hrs. Cell cultures were pelleted at 7k rpm at 4 °C for 25 minutes and were resuspended in lysis buffer containing 50 mM Tris-HCl and 300 mM NaCl at pH 8.0. Cells were lysed via sonication in the presence of 1 mg of lysozyme isolated from egg white and 2mM phenylmethanesulfonyl fluoride (PMSF) per liter of culture. Lysates were clarified by centrifugation at 15k rpm at 4 °C for 50 minutes. Clarified lysate was loaded onto HisPur™ Co^{2+}-NTA resin previously equilibrated in lysis buffer and incubated with agitation at 4 °C. The protein was washed with 10 CV of lysis buffer followed by a wash with lysis buffer containing 10mM imidazole and finally eluted using elution buffer (lysis buffer + 500 mM imidazole). Elutions were concentrated to 5 mL using an Amicon™ 10k MWCO

Ultra centrifugal filter and then dialyzed in lysis buffer without imidazole in the presence of TEV protease overnight at 4 °C. The dialyzed and post-cleaved protein was then loaded back onto the HisPur™ Co^{2+}-NTA affinity resin and eluted with several washes of lysis buffer. Elutions were further purified using a Superdex 75 preparative grade column on an AKTA-FPLC system and concentrated again prior to storage and use. Samples from pre- and post-TEV digestion of RsgI9 CRE were separated by SDS-PAGE.

NMR spectroscopy and structure calculations.

CRE was uniformly isotopically ^{13}C- and ^{15}N-labeled using the same expression and protein purification protocol as described above, but with expression induced after transferring to minimal M9 media supplemented with $^{15}NH_4Cl$ and [$^{13}C_6$] glucose. Protein samples for NMR were exchanged into NMR buffer (50mM sodium phosphate, 200 mM NaCl, 0.03% sodium azide, pH 6.0). NMR experiments were performed at 298 K on Bruker Avance III HD 600 MHz and Bruker Avance NEO 800 MHz spectrometers equipped with triple resonance cryogenic probes. Backbone and side chain atom assignments of CRE were determined via the following experiments: ^{15}N-HSQC, ^{13}C-HSQC, ^{15}N-TOCSY, HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB, HCCH-COSY, HCCH-TOCSY, HNHA, and HN(CO)CACB [28]. Data were processed using NMRPipe[54] and analyzed using XIPP NMR software [55]. Initial assignment of cross-peaks within the NOESY spectra were assigned automatically using UNIO [56] which were verified and added to manually in XIPP. $^{15}N\{^1H\}$ heteronuclear NOE data was collected in duplicate and analyzed in NMRFAM-Sparky [57]. Predicted secondary structure, RCI-S^2 values and dihedral angle restraints for structure calculations were generated using TALOS-N [58]. ^{15}N- and ^{13}C-edited NOESY spectra were acquired using a 120 ms mixing time and used to define NOE distance restraints for structural calculations made using XPLOR-NIH v3.6 [59, 60]. The final round of structural calculations included 200 structures generated, 40 of which had zero NOE violations greater than 0.5 Å or dihedral angle violations greater than 5°. A final ensemble of 20 structures based on lowest overall energy were validated via PROCHECK-NMR and selected to represent the structure of CRE, which were then deposited into the PDB (accession: 8U9O) [32]. Cartoon and ensemble representations of CRE and the AlphaFold2 model of RsgI9$^{177-350}$ were generated using MOLMOL and PyMOL [61, 62]. All backbone alignment comparisons (atoms N, Cα, C', and O) with RsgI9 CRE were done using the 'align' function in PyMOL with no pruning [62].

AlphaFold2 Modeling and MD Simulations.

A full-length, uncleaved model of RsgI9 CRE$^{177-350}$ was generated using AlphaFold2 [34] (locally installed). 5 models were generated in structure determination. The highest ranked structure based on pLDDT score was used for further MD simulations.

GROMACS version 2021 was used for explicit-solvent molecular dynamics simulations of the uncleaved AlphaFold2 model of RsgI9 CRE (G177-L350) using the CHARMM36 all-atom force field [37]. Structures were solvated in a cubic box and embedded with an appropriate amount of Na+ and Cl- counterions for an electroneutral system. Energy minimization and equilibration of the structures were completed using a steepest descent method and subsequent NVT and NPT equilibration phases lasting 100 ps each. 200 ns

simulations were performed on the equilibrated and energy minimized structures at constant temperature and pressure (300.0 K and 1.0 atm) with a time step of 2.0 fs. Global backbone RMSF over the course of the simulation reached convergence after 30 ns. The trajectory file was analyzed using VMD to visualize long-lived water molecules and changes in cleavage site side chain rotamers [63]. Fluctuations in dihedral angles over the course of the simulation were extracted from the trajectory file and visualized in XMGRACE software [64]. Hydrogen bond site occupancies were calculated using the HBonds plugin, v1.2 in VMD with a distance and angle cutoffs of 3.4 Å and 35° respectively.

Tandem mass spectrometry and MALDI-TOF.

Full length CRE was SEC purified and buffer-exchanged into 100 mM ammonium bicarbonate, pH 8.0. A tryptic digest was performed at a ratio of 50:1 RsgI9:trypsin, overnight, using sequencing grade trypsin (Promega) before desalting using C18 resin (Empore) in the format of StageTips [65]. The resulting peptides were injected onto an Exploris Orbitrap 480 (Thermo Fisher) running the following mobile phase gradient: 3% B from 0 to 7 minutes, 3% to 35% B from 7 to 40 minutes, 35% to 80% B from 40 to 45 minutes, 80% B to 50 minutes, and 80% to 3% B from 50 to 52 minutes. Data-dependent acquisition was implemented, and the resulting data was analyzed using Max Quant with the Andromeda search engine, default parameters. MS1 chromatograms were extracted using FreeStyle 1.8 (Thermo Fisher) and plotted and centered at their apices. Purified protein samples were analyzed by MALDI-TOF by first mixing with matrix solution (70% Acetonitrile, 30% ddH2O, 0.1% TFA, saturated α-cyano-4-hydroxycinnamic acid) and analyzed with an Applied Biosystems Voyager-DE STR MALDI-TOF in linear mode.

ACKNOWLEDGEMENTS

We thank members of the labs of R.T.C. and Dr. Robert Gunsalus for their support, insightful discussion, and advice, with particular thanks to Christine Minor. This research was supported by grants from the National Institute of Health (NIH) (grants S10OD025073 and S10OD016336 for partial support of the NMR core facilities, as well as NIH grant R01-AI052217 (R.T.C)). A.K.G. acknowledges support from the Whitcome Pre-doctoral Fellowship in Molecular Biology (UCLA) and the National Institute of Dental and Craniofacial Research (T90 DE030860).

FUNDING INFORMATION

This material is based upon work supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research program under Award Number DE-FC02-02ER63421, which also provides support for resources at the UCLA-DOE Macromolecular NMR Core. We also acknowledge NIH equipment grants S10OD025073 and S10OD016336 for partial support of the NMR core facilities, as well as NIH grant R01-AI052217 (R.T.C).

DATA AVAILABILITY

The coordinates and chemical shift information of the RsgI9 CRE solution structure have been deposited to the Protein Data Bank under the accession code 8U9O and the Biological Magnetic Resonance Bank under the entry ID 52129.

REFERENCES

1. IEA, World Energy Outlook 2022, IEA, Editor. 2022, International Energy Agency (IEA): Paris.

2. Bar-On YM, Phillips R, and Milo R, The biomass distribution on Earth. Proceedings of the National Academy of Sciences of the United States of America, 2018. 115(25): p. 6506–6511. [PubMed: 29784790]

3. Nassar Hussein N., a WIME, El-Gendy Nour Sh, Sustainable ecofriendly recruitment of bioethanol fermentation lignocellulosic spent waste biomass for the safe reuse and discharge of petroleum production produced water via biosorption and solid biofuel production. Journal of Hazardous Materials, 2022. 422(15): p. 126845. [PubMed: 34418833]

4. Johnson EA, et al. , Saccharification of Complex Cellulosic Substrates by the Cellulase System from Clostridium thermocellum. Applied and Environmental Microbiology, 1982. 43(5): p. 1125–1132. [PubMed: 16346009]

5. Seo H, et al. , Rewiring metabolism of Clostridium thermocellum for consolidated bioprocessing of lignocellulosic biomass poplar to produce short-chain esters. Bioresource Technology, 2023. 384: p. 129263. [PubMed: 37271458]

6. Lu Y, Zhang Y-HP, and Lynd LR, Enzyme-microbe synergy during cellulose hydrolysis by Clostridium thermocellum. Proceedings of the National Academy of Sciences of the United States of America, 2006. 103(44): p. 16165–16169. [PubMed: 17060624]

7. Alves VD, Fontes C, and Bule P, Cellulosomes: Highly Efficient Cellulolytic Complexes. Subcell Biochem, 2021. 96: p. 323–354. [PubMed: 33252735]

8. Artzi L, Bayer EA, and Morais S, Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. Nat Rev Microbiol, 2017. 15(2): p. 83–95. [PubMed: 27941816]

9. Singh N, et al. , Enzyme systems of thermophilic anaerobic bacteria for lignocellulosic biomass conversion. Int J Biol Macromol, 2021. 168: p. 572–590. [PubMed: 33309672]

10. Smith SP, Bayer EA, and Czjzek M, Continually emerging mechanistic complexity of the multi-enzyme cellulosome complex. Curr Opin Struct Biol, 2017. 44: p. 151–160. [PubMed: 28390861]

11. Leibovitz E, et al. , Characterization and subcellular localization of the Clostridium thermocellum scaffoldin dockerin binding protein SdbA. Journal of Bacteriology, 1997. 179(8): p. 2519–2523. [PubMed: 9098047]

12. Doi RH, et al. , Cellulosomes from mesophilic bacteria. Journal of Bacteriology, 2003. 185(20): p. 5907–5914. [PubMed: 14526000]

13. Dassa B, et al. , Pan-Cellulosomics of Mesophilic Clostridia: Variations on a Theme. Microorganisms, 2017. 5(4).

14. Dassa B, et al. , Genome-wide analysis of acetivibrio cellulolyticus provides a blueprint of an elaborate cellulosome system. BMC genomics, 2012. 13: p. 210. [PubMed: 22646801]

15. Artzi L, et al. , Clostridium clariflavum: Key Cellulosome Players Are Revealed by Proteomic Analysis. mBio, 2015. 6(3): p. e00411–00415. [PubMed: 25991683]

16. Nataf Y, et al. , Clostridium thermocellum cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors. Proceedings of the National Academy of Sciences of the United States of America, 2010. 107(43): p. 18646–18651. [PubMed: 20937888]

17. Kahel-Raifer H, et al. , The unique set of putative membrane-associated anti-sigma factors in Clostridium thermocellum suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation. FEMS microbiology letters, 2010. 308(1): p. 84–93. [PubMed: 20487018]

18. Sand A, et al. , Three cellulosomal xylanase genes in Clostridium thermocellum are regulated by both vegetative SigA (sigma(A)) and alternative SigI6 (sigma(I6)) factors. FEBS Lett, 2015. 589(20 Pt B): p. 3133–40. [PubMed: 26320414]

19. Munoz-Gutierrez I, et al. , Decoding Biomass-Sensing Regulons of Clostridium thermocellum Alternative Sigma-I Factors in a Heterologous Bacillus subtilis Host System. PLoS One, 2016. 11(1): p. e0146316. [PubMed: 26731480]

20. Grinberg IR, et al. , Distinctive ligand-binding specificities of tandem PA14 biomass-sensory elements from Clostridium thermocellum and Clostridium clariflavum. Proteins, 2019. 87(11): p. 917–930. [PubMed: 31162722]

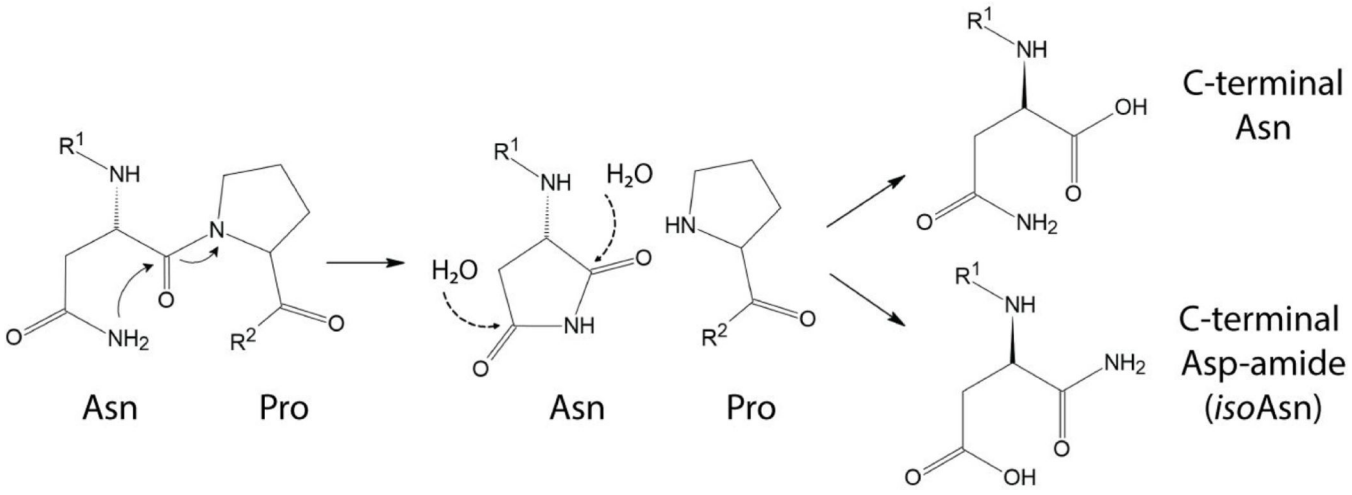
21. Bahari L, et al. , Glycoside hydrolases as components of putative carbohydrate biosensor proteins in Clostridium thermocellum. Journal of Industrial Microbiology & Biotechnology, 2011. 38(7): p. 825–832. [PubMed: 20820855]

22. Mahoney BJ, et al. , The structure of the Clostridium thermocellum RsgI9 ectodomain provides insight into the mechanism of biomass sensing. Proteins, 2022. 90(7): p. 1457–1467. [PubMed: 35194841]
23. Marcos-Torres FJ, et al. , Mechanisms of Action of Non-Canonical ECF Sigma Factors. Int J Mol Sci, 2022. 23(7).
24. Asai K, Anti-sigma factor-mediated cell surface stress responses in Bacillus subtilis. Genes & Genetic Systems, 2018. 92(5): p. 223–234. [PubMed: 29343670]
25. Brunet YR, et al. , Intrinsically disordered protein regions are required for cell wall homeostasis in Bacillus subtilis. Genes & Development, 2022. 36(17–18): p. 970–984.
26. Brogan AP, et al. , Bacterial SEAL domains undergo autoproteolysis and function in regulated intramembrane proteolysis. Proceedings of the National Academy of Sciences of the United States of America, 2023. 120(40): p. e2310862120. [PubMed: 37756332]
27. Chen C, et al. , Essential autoproteolysis of bacterial anti-σ factor RsgI for transmembrane signal transduction. Science Advances, 2023. 9(27): p. eadg4846.
28. Protein NMR spectroscopy: principles and practice. 2. ed. 2007, Amsterdam Heidelberg: Elsevier Academic Press. 885.
29. Shen Y, et al. , TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. Journal of biomolecular NMR, 2009. 44(4): p. 213–223. [PubMed: 19548092]
30. Kay LE, Torchia DA, and Bax A, Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. Biochemistry, 1989. 28(23): p. 8972–9. [PubMed: 2690953]
31. Stetz MA, et al. , Characterization of Internal Protein Dynamics and Conformational Entropy by NMR Relaxation. Methods Enzymol, 2019. 615: p. 237–284. [PubMed: 30638531]
32. Laskowski RA, et al. , AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. Journal of biomolecular NMR, 1996. 8(4): p. 477–486. [PubMed: 9008363]
33. Madeira F, et al. , Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research, 2022. 50(W1): p. W276–W279. [PubMed: 35412617]
34. Jumper J, et al. , Highly accurate protein structure prediction with AlphaFold. Nature, 2021. 596(7873): p. 583–589. [PubMed: 34265844]
35. Chen VB, et al. , MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica. Section D, Biological Crystallography, 2010. 66(Pt 1): p. 12–21. [PubMed: 20057044]
36. Prisant MG, et al. , New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink “waters,” and NGL Viewer to recapture online 3D graphics. Protein Sci, 2020. 29(1): p. 315–329. [PubMed: 31724275]
37. Huang J and MacKerell AD, CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. Journal of Computational Chemistry, 2013. 34(25): p. 2135–2145. [PubMed: 23832629]
38. Guan C, et al. , Activation of glycosylasparaginase. Formation of active N-terminal threonine by intramolecular autoproteolysis. The Journal of Biological Chemistry, 1996. 271(3): p. 1732–1737. [PubMed: 8576176]
39. Johansson DG, et al. , Protein autoproteolysis: conformational strain linked to the rate of peptide cleavage by the pH dependence of the N --> O acyl shift reaction. J Am Chem Soc, 2009. 131(27): p. 9475–7. [PubMed: 19534521]
40. Voorter CE, et al. , Spontaneous peptide bond cleavage in aging alpha-crystallin through a succinimide intermediate. The Journal of Biological Chemistry, 1988. 263(35): p. 19020–19023. [PubMed: 3198609]
41. Boon L, et al. , Protease propeptide structures, mechanisms of activation, and functions. Critical Reviews in Biochemistry and Molecular Biology, 2020. 55(2): p. 111–165. [PubMed: 32290726]
42. Kemp MT, Lewandowski EM, and Chen Y, Low barrier hydrogen bonds in protein structure and function. Biochimica Et Biophysica Acta. Proteins and Proteomics, 2021. 1869(1): p. 140557. [PubMed: 33148530]

43. Perler FB, Xu MQ, and Paulus H, Protein splicing and autoproteolysis mechanisms. Current Opinion in Chemical Biology, 1997. 1(3): p. 292–299. [PubMed: 9667864]
44. Bastiaansen KC, et al. , Self-cleavage of the Pseudomonas aeruginosa Cell-surface Signaling Anti-sigma Factor FoxR Occurs through an N-O Acyl Rearrangement. The Journal of Biological Chemistry, 2015. 290(19): p. 12237–12246. [PubMed: 25809487]
45. Araç D, et al. , A novel evolutionarily conserved domain of cell-adhesion GPCRs mediates autoproteolysis. The EMBO journal, 2012. 31(6): p. 1364–1378. [PubMed: 22333914]
46. Hsu M-F, et al. , Mechanism of the maturation process of SARS-CoV 3CL protease. The Journal of Biological Chemistry, 2005. 280(35): p. 31257–31266. [PubMed: 15788388]
47. Ball LE, et al. , Post-translational modifications of aquaporin 0 (AQP0) in the normal human lens: spatial and temporal occurrence. Biochemistry, 2004. 43(30): p. 9856–9865. [PubMed: 15274640]
48. Ferris HU, et al. , FlhB regulates ordered export of flagellar components via autocleavage mechanism. The Journal of Biological Chemistry, 2005. 280(50): p. 41236–41242. [PubMed: 16246842]
49. Monjarás Feria JV, et al. , Role of autocleavage in the function of a type III secretion specificity switch protein in Salmonella enterica serovar Typhimurium. mBio, 2015. 6(5): p. e01459–01415. [PubMed: 26463164]
50. Kato K, et al. , Computational Analysis of the Mechanism of Nonenzymatic Peptide Bond Cleavage at the C-Terminal Side of an Asparagine Residue. ACS omega, 2021. 6(44): p. 30078–30084. [PubMed: 34778679]
51. Robinson NE and Robinson AB, Molecular clocks. Proceedings of the National Academy of Sciences of the United States of America, 2001. 98(3): p. 944–949. [PubMed: 11158575]
52. Geiger T. and Clarke S, Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. The Journal of Biological Chemistry, 1987. 262(2): p. 785–794. [PubMed: 3805008]
53. Tarelli E. and Corran PH, Ammonia cleaves polypeptides at asparagine proline bonds. The Journal of Peptide Research: Official Journal of the American Peptide Society, 2003. 62(6): p. 245–251.
54. Delaglio F, et al. , NMRPipe: a multidimensional spectral processing system based on UNIX pipes. Journal of biomolecular NMR, 1995. 6(3): p. 277–293. [PubMed: 8520220]
55. Garrett DS, Cai M, and Clore GM, XIPP: multi-dimensional NMR analysis software. Journal of biomolecular NMR, 2020. 74(1): p. 9–25. [PubMed: 31748843]
56. Herrmann T, Güntert P, and Wüthrich K, Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. Journal of Molecular Biology, 2002. 319(1): p. 209–227. [PubMed: 12051947]
57. Lee W, Tonelli M, and Markley JL, NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics (Oxford, England), 2015. 31(8): p. 1325–1327. [PubMed: 25505092]
58. Shen Y. and Bax A, Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. Journal of biomolecular NMR, 2013. 56(3): p. 227–241. [PubMed: 23728592]
59. Schwieters CD, Bermejo GA, and Clore GM, Xplor-NIH for molecular structure determination from NMR and other data sources. Protein Science: A Publication of the Protein Society, 2018. 27(1): p. 26–40.
60. Schwieters CD, et al. , The Xplor-NIH NMR molecular structure determination package. Journal of Magnetic Resonance (San Diego, Calif.: 1997), 2003. 160(1): p. 65–73. [PubMed: 12565051]
61. Koradi R, Billeter M, and Wüthrich K, MOLMOL: a program for display and analysis of macromolecular structures. Journal of Molecular Graphics, 1996. 14(1): p. 51–55, 29–32. [PubMed: 8744573]
62. Lam WWT and Siu SWI, PyMOL mControl: Manipulating molecular visualization with mobile devices. Biochemistry and Molecular Biology Education: A Bimonthly Publication of the International Union of Biochemistry and Molecular Biology, 2017. 45(1): p. 76–83.
63. Humphrey W, Dalke A, and Schulten K, VMD: visual molecular dynamics. Journal of Molecular Graphics, 1996. 14(1): p. 33–38, 27–28. [PubMed: 8744570]

64. Turner PJ, XMGRACE. 2005, Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology: Beaverton, OR.

65. Rappsilber J, Mann M, and Ishihama Y, Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nature Protocols, 2007. 2(8): p. 1896–1906. [PubMed: 17703201]



Scheme 1.

Schematic showing the potential mechanism for the spontaneous deamidation and cleavage of the N188/P189 bond. Auto-cleavage of the peptide bond acts as the initial step leading to the formation of the succinimide intermediate. The C-terminal succinimide undergoes hydrolysis leading to the formation of either a C-terminal Asn residue or a C-terminal Asp amide (isoAsn). R1 and R2 represent the N- and C-terminal ends of the polypeptide chain, respectively.

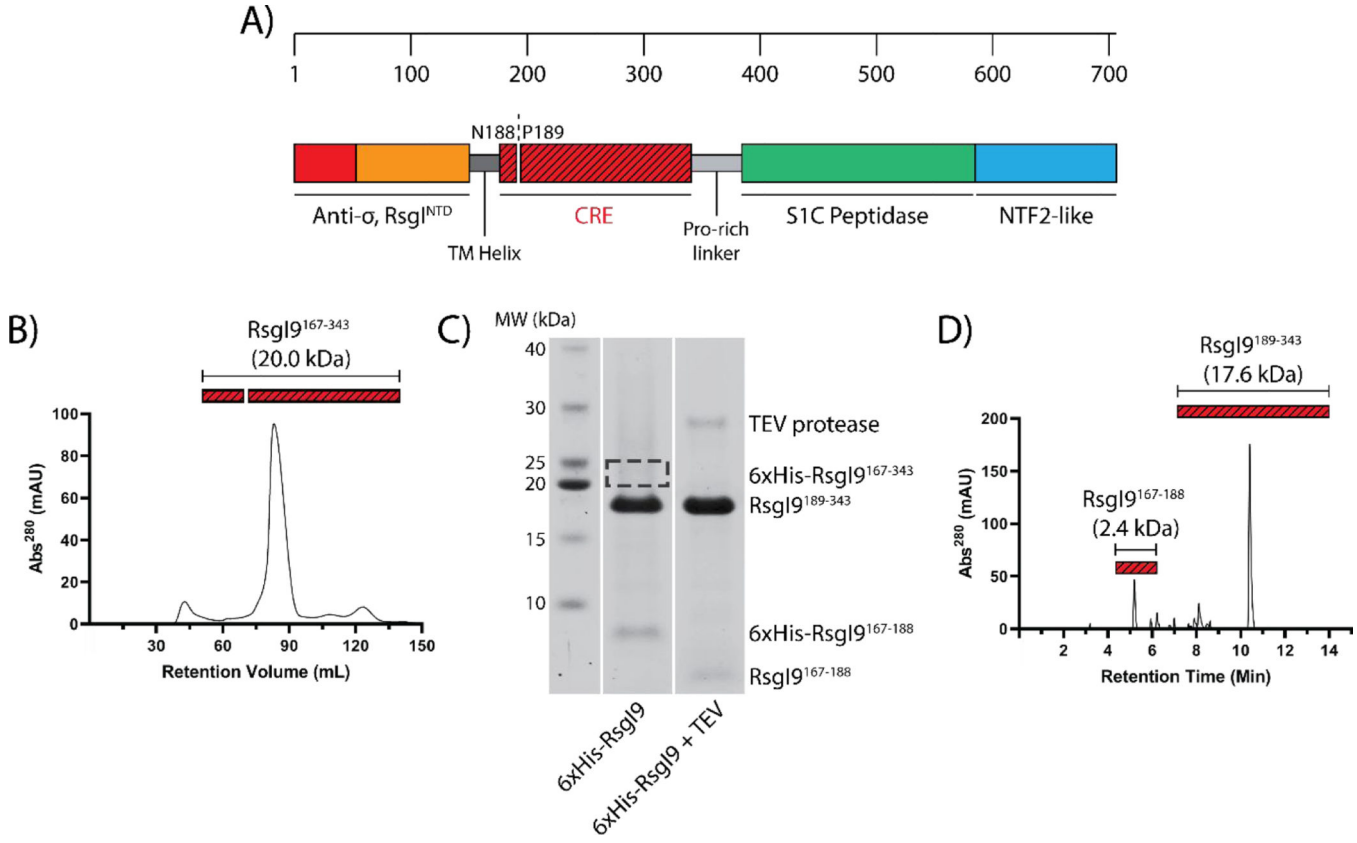


Figure 1. The RsgI9 anti-σ factor is proteolyzed at the Asn^{188}-Pro^{189} peptide bond.

(A) Schematic of the intact RsgI9 protein. It contains conserved domains that are found in other RsgI-family proteins (red). These include an intracellular anti-σ factor domain (NTD, residues 1– 55, red), a transmembrane helix (TM, dark gray) and an extracellular domain of unknown function (CRE, residues 167–343, red, striped). RsgI9 and many other RsgI proteins also contain a Pro-rich linker segment (light gray) that connects the RsgI9 CRE to S1C peptidase (mint, residues 396–578) and NTF2-like type domains (blue, residues 579–707). RsgI9 also contains a unique insertion (orange) immediately following the NTD that is in the cytoplasm. (B) SEC chromatogram of purified CRE protein, demonstrating that it elutes as a single peak. The elution volume for CRE is consistent with the expected molecular weight of the uncleaved form. (C) SDS-PAGE of His-tagged RsgI9 CRE (167–343) digested with TEV protease. A band corresponding to a fully intact, undigested 6xHis-RsgI9 (residues 167–343) is absent in the predigestion sample. (D) Reversed-phase HPLC chromatogram of purified CRE obtained from SEC. The data show that CRE consists of two polypeptides, short (residues G167 to N188, RsgI9) and long (residues 189–343, RsgI9) fragments. The identity of the polypeptides was confirmed by ESI-MS/MS.

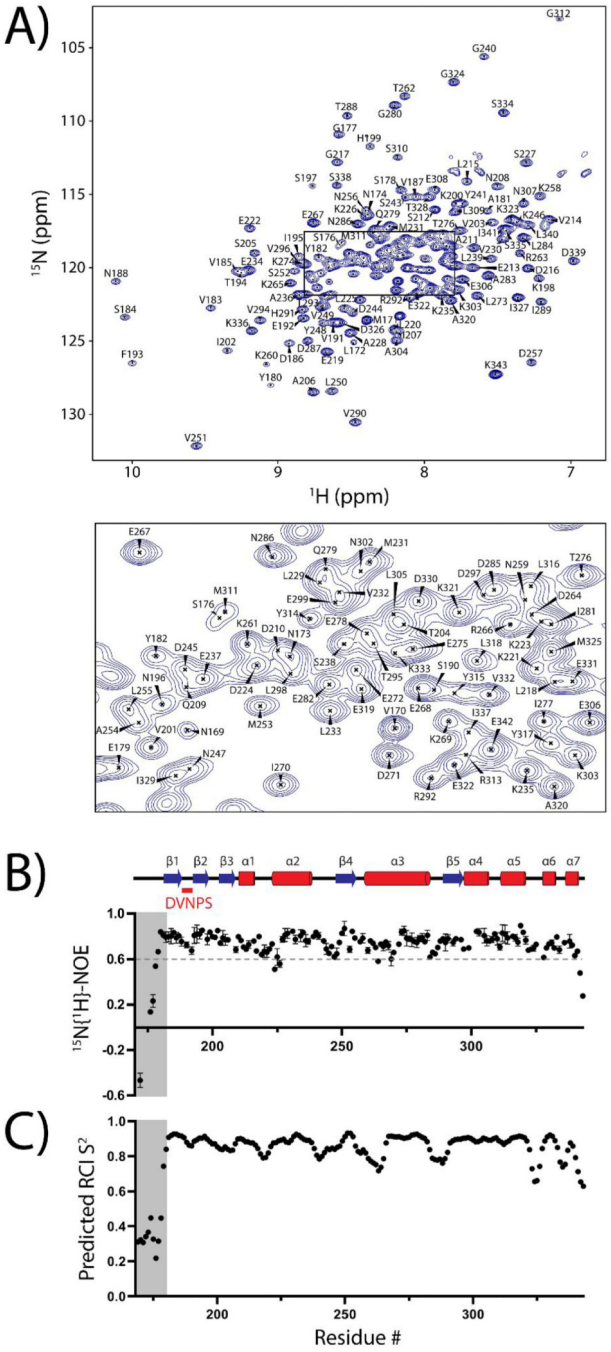


Figure 2. NMR data showing the cleaved CRE domain forms a stable structure.

(A) The ^{1}H-^{15}N HSQC spectrum of the uniformly ^{15}N-labeled RsgI9 CRE domain. Residue numbering corresponds to the full-length RsgI9 protein. (B) Plot of the steady state heteronuclear ^{15}N{^{1}H} NOE (hetNOE) values for the backbone amides in the domain plotted as a function of residue number. Error values are the maximal deviation from the average obtained from two separate measurements. A schematic of the secondary structural elements identified by NMR is depicted above. Out of a total of 176 residues containing backbone amide groups in CRE, resolved hetNOE data could be obtained for 165

residues. (C) Predicted Random Coil Index (RCI) values per residue as calculated using the program TALOS-N and chemical shift data. In panels (B) and (C) the light-gray shaded area represents the first 10 residues of the CRE protein construct and are structurally disordered.

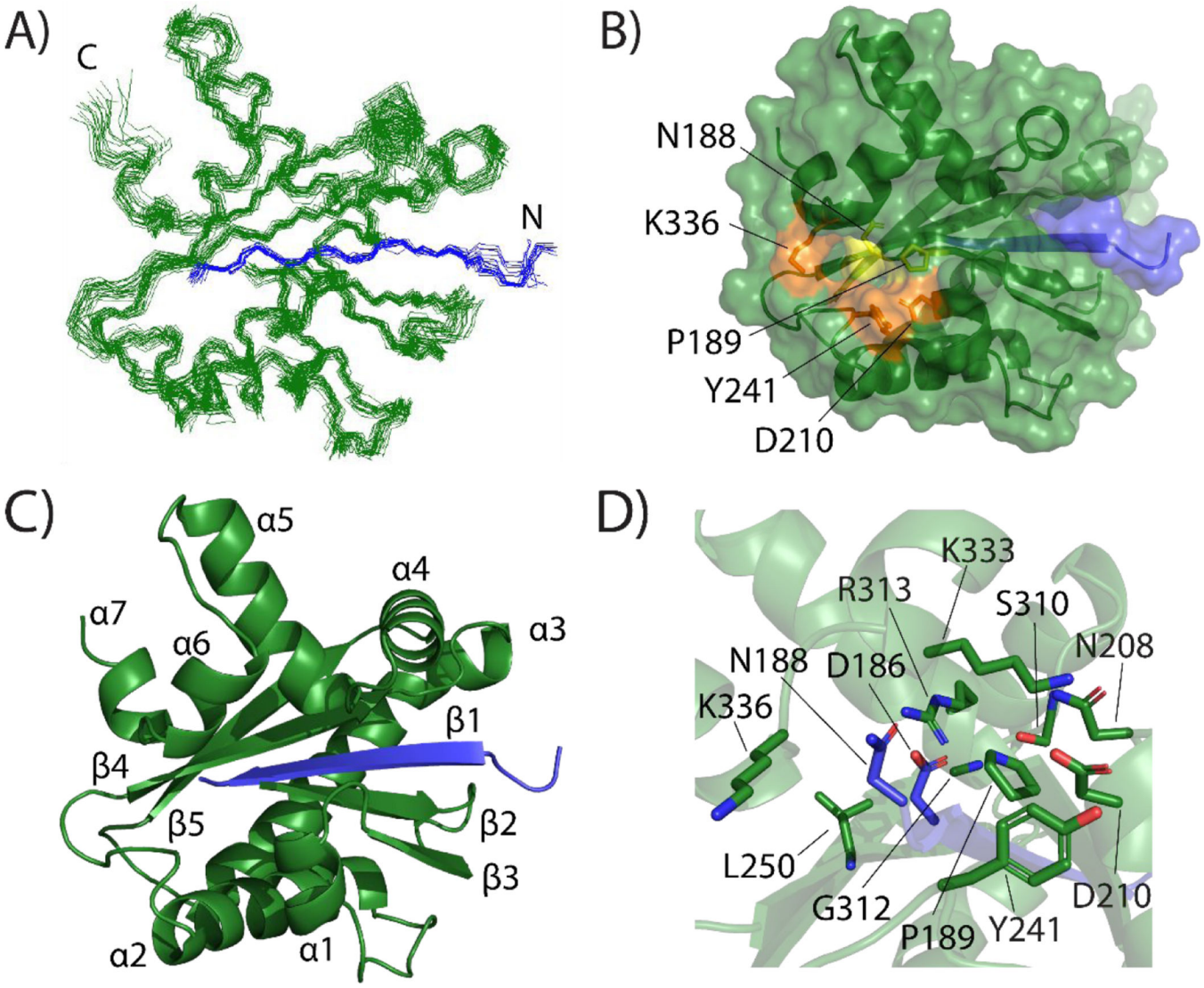


Figure 3. Solution structure of the RsgI9 CRE domain.

(A) Superposition of the 20 lowest energy structures of the RsgI9 CRE domain. Residues corresponding to the 2.4 and 17.6 kDa fragments are colored light and dark blue, respectively. Residues G167-K343 are shown with their N- and C-termini labeled. (B) Surface representation of the CRE domain structure. Cleavage site residues (yellow) and residues that partially occlude the cleavage site (orange) are colored. (C) Ribbon representation of the lowest energy structure of the CRE domain with its secondary structural elements labeled (α1–7, β1–5) (pdb accession: 8U9O) (D) Expanded view of the structure showing the cleaved peptide bond between the Asn^{188} and Pro^{189} residues. Carbon atoms are colored by fragment (G167-N188, blue and P189-K343, green). The side chains of nearby conserved residues are shown in stick format.

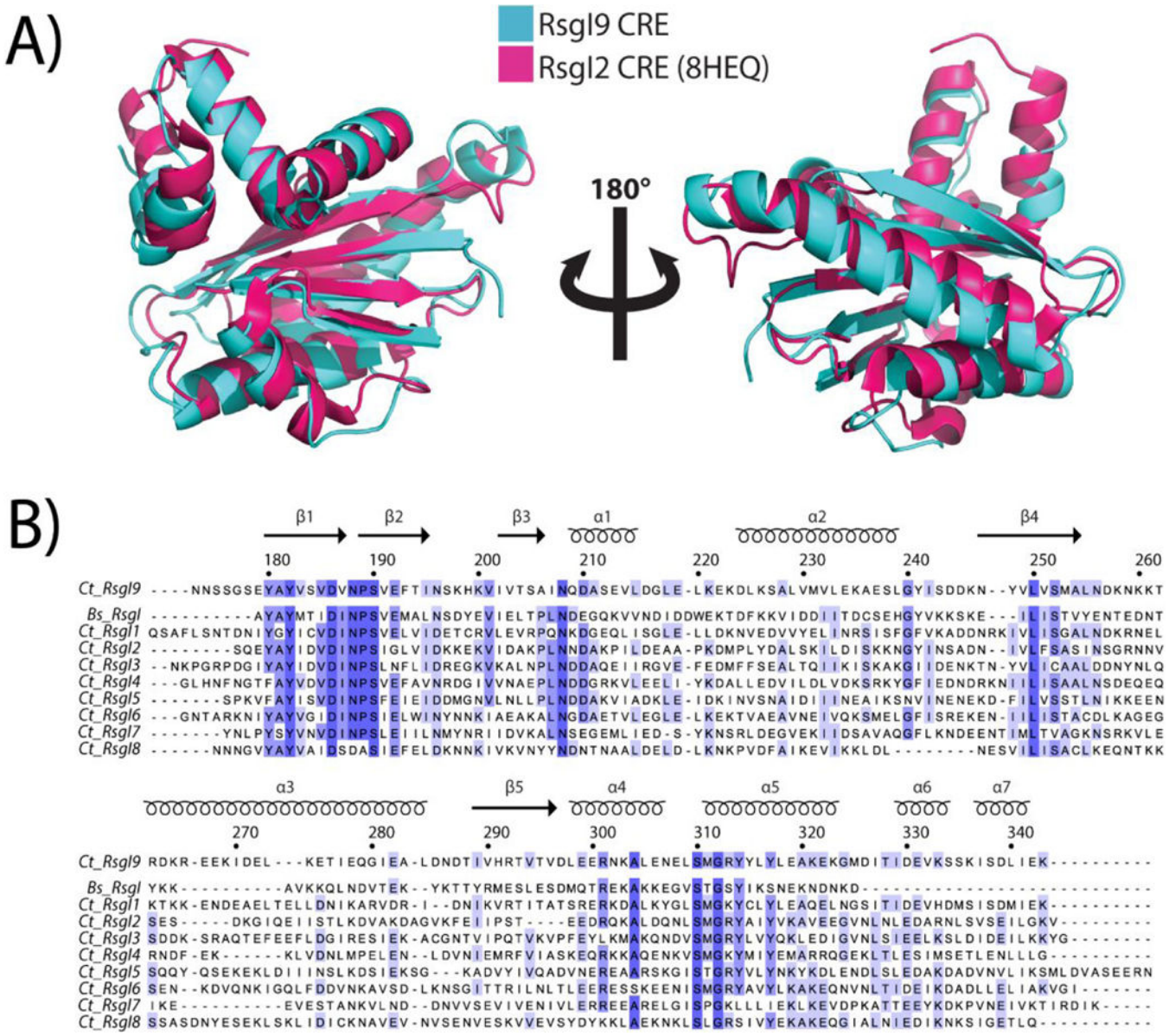


Figure 4. Structure and sequence comparisons with the RsgI9 CRE domain.

(A) Overlay of the lowest energy structure of the RsgI9 CRE domain (cyan, residues 167–343) with the structure of the CRE domain from the *C. thermocellum* RsgI2 protein (magenta, residues 89–248) (pdb: 8heq)[27]. The structures align with a backbone RMSD of 1.95 Å. (B) Sequence alignment of the CRE domain (top) with the other RsgI proteins in *C. thermocellum* (RsgI 1–8) and the RsgI anti-σ factor from *B. subtilis* (Bsub_RsgI). Residue numbers and secondary structures are shown above the sequences and correspond to the RsgI9 protein. Shaded residues are conserved (dark purple most conserved). Multiple sequence alignments for the RsgI proteins were generated using Clustal Omega [33].

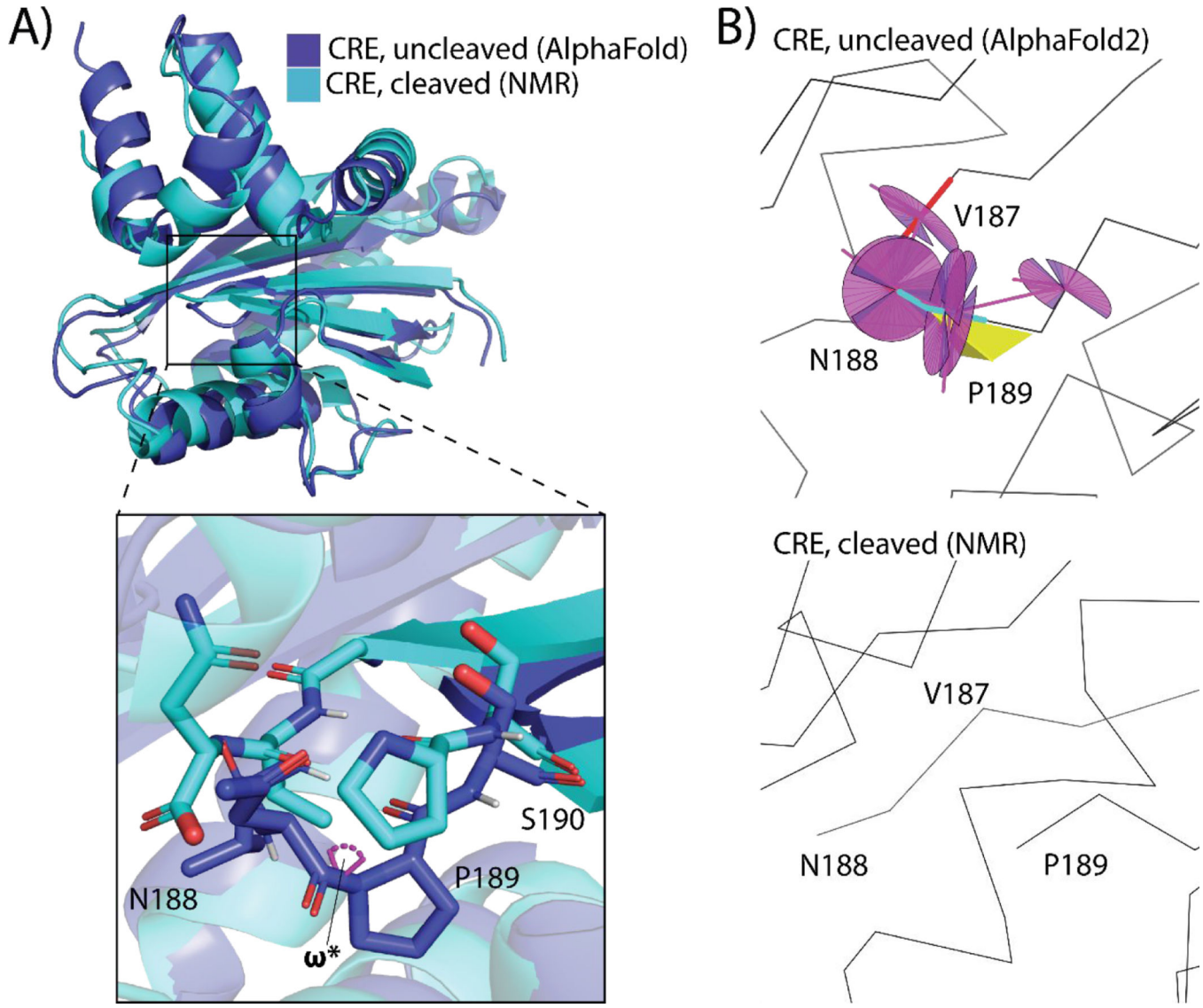


Figure 5. AlphaFold2 Comparison to solution-state structure of CRE.

(A) Overlay and zoom of the lowest overall energy structure calculated for RsgI9 CRE (167–343) and the AlphaFold2-predicted model of CRE (177–350). The solution-state NMR structure of CRE contains the cleavage site breakage between N188 and P189 whereas the predicted AlphaFold2 structure is continuous through that region. The two structures align with a backbone RMSD of 1.94 Å. The conserved cleavage site residues (N188/P189) are compared in the zoom-in window. The ω* denotes a strained, non-planar peptide angle (83.2°) for P189 in the uncleaved AlphaFold2 model. (B) MolProbity kinemage analysis of the CRE cleavage site residues for the uncleaved AlphaFold2 model and solution state structure of CRE. Backbone carbons are outlined in black. Outlier values for Ramachandran (thick cyan line), bond-length (thick red line), twisted/non-planar peptide bonds (yellow trapezoid), and C-Alpha geometry (magenta discs) are shown.

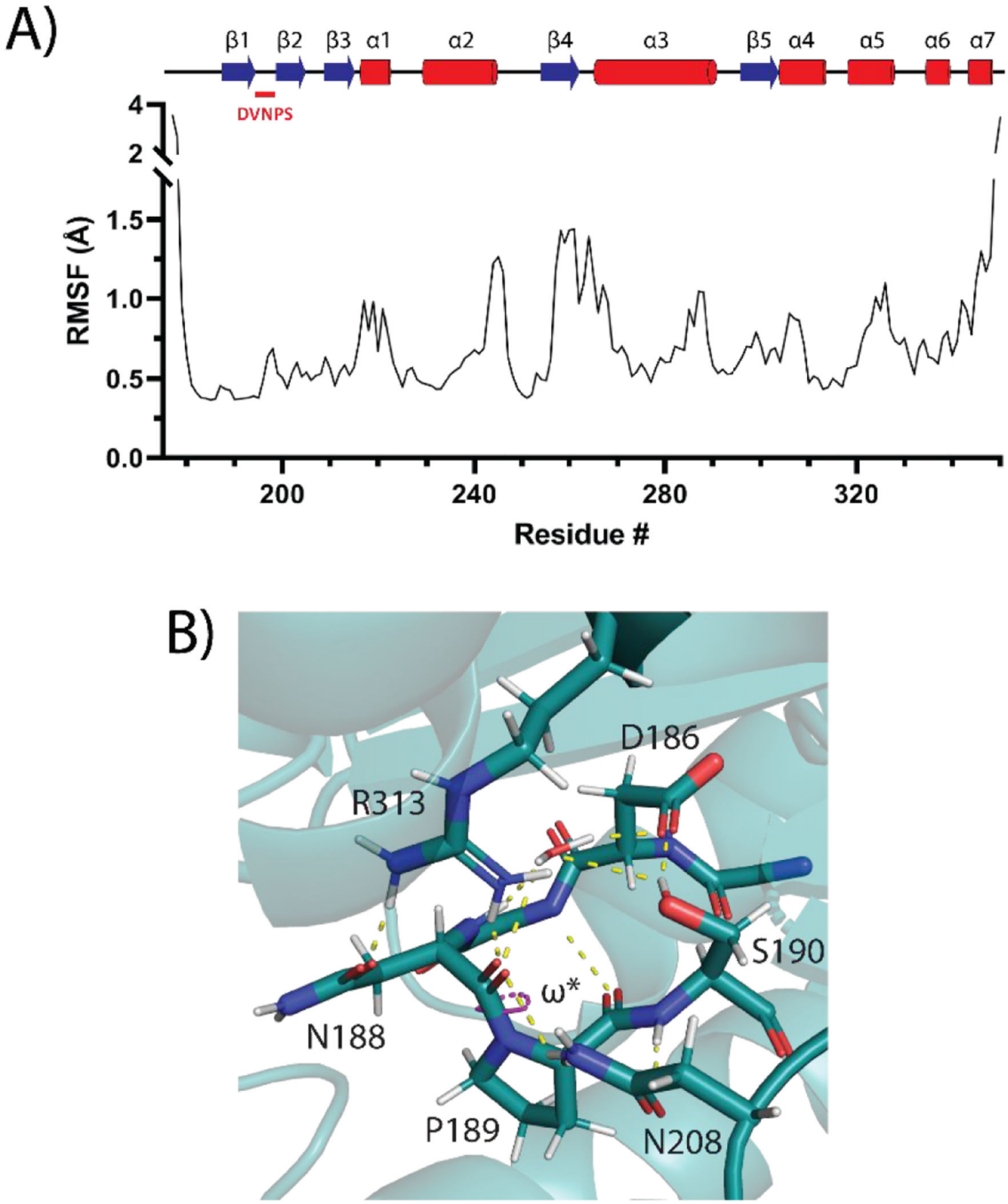


Figure 6. MD simulations of the AlphaFold2 model of CRE.

(A) Plot of the root mean square fluctuation (RMSF) differences of CRE backbone amide coordinates during the course of the 200 ns MD simulation. (B) Zoom in view of a representative frame from an uncleaved precursor $RsgI9^{CRE}$ during the course of the MD simulation. The peptide bond dihedral angle for P189 (ω*, magenta) remains predominantly in a non-planar (153.2°) configuration as stabilized by the surrounding hydrogen bond network. Backbone amide and side chain hydrogens (white) are shown to illustrate peptide

bond geometries. Heavy atoms are colored blue and red to indicate nitrogen and oxygen, respectively. Hydrogen bonds are denoted with dashed yellow lines.

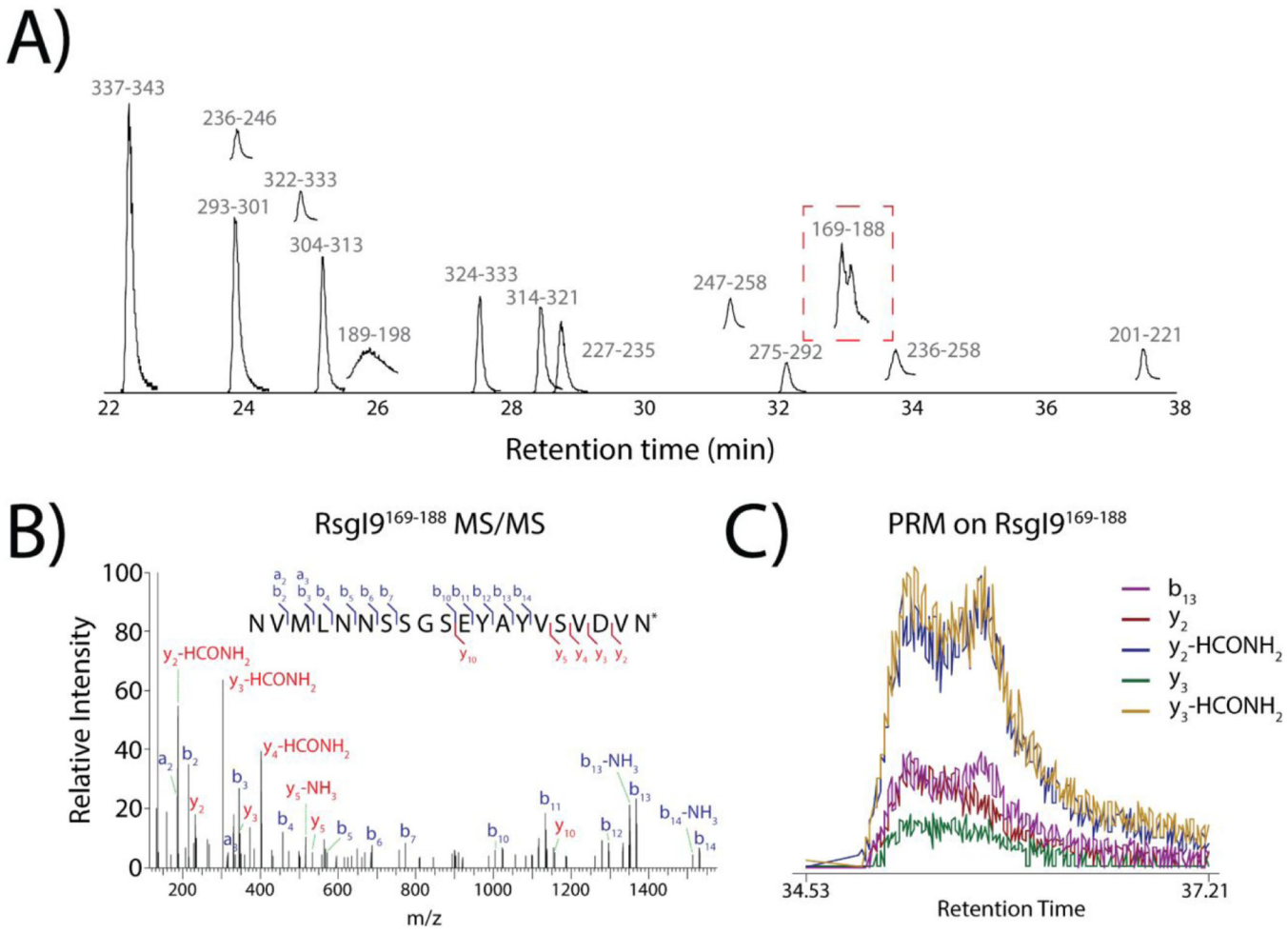


Fig 7. LC-MS/MS Chromatogram reveals bimodal peak for N-terminal CRE cleavage fragment.

(A) LC-MS extracted ion chromatograms of RsgI9 peptides. Numbers above each peak indicate the residue numbers of full-length RsgI9 CRE represented in each fragment. Those on the x-axis are to-scale and those above the x-axis are zoomed in to highlight their elution. Note peptide 169–188 is the only peptide with a bimodal elution. (B) Annotated MS2 spectra from peptide 169–188 confirm its identity. An asterisk (*) is used to denote either Asn/*iso*Asn (C) Parallel reaction monitoring (PRM) of the 169–188 product ions with an optimized elution gradient. The intensities of ions containing N22 residue (y_2, y_3, and dissociated products) as well as b_{13} are plotted, which shows the entire bimodal elution profile corresponds to elution peak CRE (169–188).

Table 1.

Structural statistics of the solution structure of CRE

	⟨SA⟩[a]
Root mean square deviation	
NOE Interproton distance restraints (Å)(1668)[b]	0.028 ± 0.002
Dihedral angles restraints (°) (301)[c]	0.550 ± 0.059
Hydrogen bond distance restraints (Å) (74)	0.027 ± 0.005
Deviation from idealized covalent geometry	
bonds (Å)	0.002 ± 0.0001
angles (°)	0.425 ± 0.011
impropers (°)	0.301 ± 0.017
PROCHECK results (%)	
most favorable region	94.8 ± 1.1
additionally allowed region	4.4 ± 1.1
generously allowed region	0.7 ± 0.4
disallowed region	0.0 ± 0.0
Coordinate Precision (Å)[d]	
Protein backbone	0.57 ± 0.07
Protein heavy atoms	1.16 ± 0.07

[a]⟨SA⟩ represents an ensemble of the 220 best structures calculated by simulated annealing. The number of terms for each restraint is given in parentheses. None of the structures exhibited distance violations greater than 0.5 Å, or dihedral angle violations greater than 5°. Residues selected for statistics and PROCHECK-NMR analysis were Y180-I337. [32]

[b]Distance restraints: 463 sequential, 261 medium (2 ≤ residue separation ≤ 4), 506 long range (>4 residues apart) and 438 intramolecular.

[c]The experimental dihedral angle restraints were as follows: 3151 φ and 151 ψ angular restraints

[d]The coordinate precision is defined as the average atomic root mean square deviation (RMSD) of the 420 individual SA structures and their mean coordinates. These values are for residues Y180-I337 of CRE. Backbone atoms refer to the N, C^{α}, and C' atoms.