TITLE

Structural, dynamic and functional characterization of a new DnaX mini intein derived from *Spirulina platensis* provides new insights into intein-mediated catalysis of protein splicing

Soumendu Boral[a], Snigdha Maiti[a], Aditya J Basak[a], Woonghee Lee[b] and Soumya De[a]*

*Corresponding author
[a]School of Bioscience, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India
[b]Department of Chemistry, University of Colorado Denver, Denver, CO 80217, USA

KEY WORDS

NMR Spectroscopy
Site-directed mutagenesis
Cleavage reaction
Hydrogen exchange

ABSTRACT

Protein splicing is a self-catalyzed post-translational modification in which an intervening protein or intein excises itself out from a precursor protein, and ligates the flanking external proteins or exteins to produce a mature protein. We report the solution NMR structure of a 136-residue DnaX mini-intein enzyme from the cyanobacterium *Spirulina platensis*. The protein adopts a well-defined globular structure, representing a canonical horseshoe-like disk-shaped fold commonly found in the HINT (hedgehog intein)-type topology. Analytical ultracentrifugation and size-exclusion chromatography confirm the monomeric nature of the protein in solution. NMR-based backbone dynamics and hydrogen exchange experiments reveal conserved motions in both fast (ps-ns) and slow (μs-ms) timescales, which appear to be a characteristic of the intein fold. In cell splicing activity of *Spl* DnaX mini intein was determined by SDS PAGE and western blotting, and the enzyme was found to be highly active. Apart from splicing reaction, the precursor protein also undergoes catalytic cleavage at the N- and C-terminus of the intein enzyme. To determine the roles of the three catalytic residues in splicing and cleavage reactions, all possible alanine mutations of these residues were generated and functionally characterized. This study revealed cooperativity between these

catalytic residues, which suppresses the N- and C-terminal cleavage reactions and thus, enhances the yield of the spliced product. Overall, this study provides detailed structural, dynamic and functional characterization of a new intein enzyme and adds to the collection of these unique enzymes that have found tremendous applications in chemical biology and biotechnology.

INTRODUCTION

Protein splicing is a unique post-translational rearrangement that was discovered in the early 1990s.[1–3] This process involves self-excision of an internal protein sequence, called an intein (intervening protein), from a precursor protein, resulting in the ligation of the two flanking N- and C-exteins (external proteins) into a single polypeptide chain via a native peptide bond formation between them.[4] Inteins have been identified in a number of unicellular organisms and occur in all three domains of life: bacteria, eukarya and archaea.[5] Since their discovery, intein enzymes have found numerous applications in chemical biology, protein engineering, biomedicine and biotechnology.[6,7] Indeed, these enzymes have been termed as 'Nature's gift to protein chemists'.[8] Few examples of their applications are expressed protein ligation (EPL),[9,10] conditional protein splicing (CPS),[11,12] self-cleavage of affinity-tag for protein purification, site-specific chemical modification,[13] segmental isotope labeling [14,15] and protein cyclization.[16] Till date, more than 600 intein genes have been found and are deposited into the New England Biolabs Intein Database, InBase,[17] but only a few have been thoroughly characterized. Since inteins have a wide range of applications in protein chemistry, it is important to characterize more intein sequences that may find novel applications.

Protein splicing is an intramolecular reaction. Inteins are single turn-over enzymes and require neither cofactors nor energy sources. Based on their sequences and splicing mechanisms, inteins are grouped into three classes.[18] For class 1 inteins, splicing is initiated by the N-terminal Cys or Ser sidechain of the intein, which attacks the previous peptide bond resulting in a thioester/ester linkage at the N-terminal splice junction (Figure 1). In the second step, the N-extein is then transferred to the side chain of the first C-extein residue (Cys, Ser or Thr) by *trans*-(thio)esterification, resulting in a branched (thio)ester intermediate. In the third step, the sidechain of the C-terminal asparagine attacks the following peptide bond resulting in peptide bond cleavage and separates the ester-linked exteins from the intein. Finally, the ester linkage rearranges to an amide linkage and the succinimide ring at the intein's C-terminal end hydrolyzes to form Asn.[19] This splicing reaction is catalyzed by a set of conserved active site residues of the intein enzyme.

However, both class 2 and class 3 inteins lack the N-terminal Cys or Ser nucleophile, thus they are unable to form the linear thio-ester intermediate (Figure 1). In case of class 2 inteins,[20,21] the sidechain of the first C-extein residue directly attacks the N-terminal scissile bond, thereby omitting both the acyl rearrangement and trans-esterification reaction and

resulting in the formation of the branched intermediate as in class 1 inteins. Class 3 inteins contain a conserved Trp-Cys-Thr (WCT) triplet, whose Cys sidechain attacks the N-terminal scissile amide bond, resulting in a branched intermediate.[18,22] In the second step, the N-extein is transferred from the sidechain of this Cys to the sidechain of first C-extein residue to form another branched intermediate, similar to class 1 and 2 inteins. Once the standard branched intermediate is formed, both the class 2 and class 3 inteins follow the same class 1 splicing pathway.[6]

In this study, we have carried out structural and functional characterization of a new intein sequence using solution NMR spectroscopy and in cell splicing assay, respectively. The intein was derived from gamma and tau subunits of DNA polymerase III (DnaX) gene of a cyanobacterium *Spirulina platensis*, abbreviated as *Spl* DnaX intein.[17] This intein contains 136 residues and is classified as a mini-intein due to the absence of a homing endonuclease domain. We have solved the structure of *Spl* DnaX mini-intein by solution NMR spectroscopy. Similar to other intein structures,[23–26] the global fold of *Spl* DnaX mini-intein comprises of several β-strands forming a horseshoe-like disk shaped fold commonly found in HINT (hedgehog/intein) domain superfamily. Furthermore, NMR-based hydrogen exchange experiments and fast ps-ns backbone amide dynamics reveal that the core of the intein enzyme is highly rigid. The splicing reaction, within the bacterial cell, was monitored by gel electrophoresis of the whole cell lysate and the splicing and cleavage products were detected by western blotting analysis. Based on the solved structure, the catalytically important residues were identified and their functional role in the catalytic mechanism was determined by site-directed mutagenesis. Overall, in this study, we show that the *Spl* DnaX mini-intein has a well-folded and rigid three-dimensional structure, is highly active and follows the class 1 splicing mechanism.

MATERIALS AND METHODS

Cloning, protein expression and purification

The full-length (138 aa) wild type *Spl* DnaX mini intein from *Spirulina platensis* was cloned into pET28a(+) vector. The plasmid was transformed into *Escherichia coli* BL21(λDE3) strain. For ^{15}N labeled protein sample, 1 L of M9 minimal media was supplemented with 1g of $^{15}NH_4Cl$ and 5g of $^{12}C_6$-glucose as the sole nitrogen and carbon source, respectively. The culture was grown at 37°C till the cell density reached OD_{600} of ~1.0. After induction with 0.5 mM IPTG for 5 hours at 37°C, cells were harvested and lysed by sonication in a lysis buffer (100 mM Tris-Cl, 200 mM NaCl, 10 mM Imidazole, pH 8.2). The wildtype intein cleaved off the N-terminal $(His)_6$-tag, thus hampering affinity purification. Hence, after cell lysis, the lysate was centrifuged at 16,000g for 50 minutes to precipitate the cell debris. The supernatant was heated to 70°C for 15 minutes and centrifuged to precipitate unfolded protein aggregates. The resulting clarified supernatant was filtered and passed through an anion and cation exchange resins, respectively. Finally, size exclusion chromatography was performed to obtain highly pure intein protein.

For structural studies, a catalytically dead mutant intein was designed by mutating Cys1 and Cys(+1) to Ala. The resulting construct Intein$^{C1A/C(+1)A}$ (Table 1) was over-expressed in 1 litre of M9 minimal media supplemented with 1g of $^{15}NH_4Cl$ and 3g of $^{13}C_6$-glucose and purified using a Ni-NTA column. The affinity tag was removed by the Thrombin CleanCleave Kit from Sigma-Aldrich. The protein was exchanged into the final buffer (20 mM sodium phosphate, 50 mM NaCl, and pH 6.5). Protein concentration was determined by UV absorption using predicted molar absorptivity (ε_{280}). The proteins were 0.6-0.7 mM with 7% lock D_2O. For the long-term stability of the proteins, 0.8 mM PMSF, 2 µl protease inhibitor cocktail and 0.04% NaN_3 were also added to the final sample.

Analytical size exclusion chromatography (SEC)

Purified Intein^{WT} and Intein$^{C1A/C(+1)A}$ in a buffer containing 20 mM Tris (pH 7.5), 50 mM NaCl were subjected to pre-packed Biorad ENrich^{TM} SEC 70 size exclusion column. Analytical SEC profiles were compared to probe any hydrodynamic shape difference between the two protein constructs.

Analytical ultracentrifugation (AUC)

Sedimentation velocity analytical ultracentrifugation (SV-AUC) experiment was performed on Intein^{WT} and the double mutant Intein$^{C1A/C(+1)A}$ using a ProteomeLab XL-1 analytical ultracentrifuge (Beckman Coulter). Epon double-sector centrepieces were filled with 420 µl of the sample buffer (20 mM Tris, 50 mM NaCl, pH 7.5) and 400 µl of intein protein having an A_{280} value of 0.4–0.6, respectively in the two sectors. The samples were centrifuged at 40,000 rpm at 20°C. Frames were collected until sedimentation was complete. Absorbance scans were taken at intervals of 3 minutes. Sample buffer density, viscosity and partial specific volume of the proteins were calculated by SEDNTERP,[27] data analysis was performed by SEDFIT using c(S) distribution analysis,[28] and figures were prepared using GUSSI.[29]

NMR experiments and structure calculations

NMR experiments were performed in a 600 MHz Bruker Avance III spectrometer equipped with a triple resonance cryogenic probe head. All studies were done at 303 K unless otherwise stated. For the backbone and sidechain assignment, the ^{15}N, ^{13}C double-labeled Intein$^{C1A/C(+1)A}$ protein sample was used. Backbone resonances were manually assigned in NMRFAM-SPARKY 1.470 [30] using 2D ^{15}N-^{1}H HSQC and 3D NMR experiments such as CBCA(CO)NH, CBCANH, HNCO, HN(CA)CO and HNCA. Sidechains were assigned using HBHA(CO)NH, H(CC)(CO)NH, (H)CC(CO)NH, HCCH-TOCSY and ^{15}N-TOCSY experiments.[31] These spectra were processed and analyzed using NMRPipe[32] and Sparky,[33] respectively. The secondary structure propensity of this protein was calculated from the backbone chemical shifts ($^{1}H^{N}$, ^{15}N, $^{13}C_\alpha$, $^{13}C_\beta$ and ^{13}CO) using the program MICS.[34] For the structure calculation, a ^{15}N-edited NOESY spectrum with a mixing time of 110 ms and a sensitive enhanced ^{13}C-edited NOESY spectrum with a mixing time of 110 ms were recorded.

Chemical shift-based torsion angle restraints and NOE based distance restraints were used to calculate the three-dimensional structure of the intein. Initial steps of NOE assignment and fold calculation were performed with PONDEROSA.[35] Xplor-NIH-based calculations[36] implemented in AUDANA[37] were used for the subsequent steps in the PONDEROSA-C/S package.[38] Structure calculations were carried out by submitting jobs to the Ponderosa web server.[39] The final structures were calculated with explicit water refinement. Secondary structural boundaries were determined using DSSP[40] and structures were generated with PyMOL (PyMOL molecular graphics system, version 2.0.4, Schrodinger, LLC, New York, NY, USA). Statistics for the 15 lowest energy conformers are summarized in Table 2.

Backbone amide ^{15}N relaxation

Amide ^{15}N relaxation (R_1, R_2 and steady-state heteronuclear NOE) experiments were performed at 303K for Intein$^{C1A/C(+1)A}$ protein as described previously.[41,42] Briefly, R_1 (50, 100, 200, 300, 400, 500, 750, 1000 and 1200 ms) and R_2 (20, 40, 60, 80, 100, 120, 140, 160, 180 and 200 ms) spectra were collected in random order to minimize any systematic error. The peak intensities of each residue were fit to an exponential decay function [$I_t = I_0*\exp(-t*R_i)$], where I_t is the peak intensity, t is the relaxation delay, I_0 is the initial intensity and R_i is either R_1 or R_2. Errors in the calculated rate constants were determined by Monte-Carlo simulation. The ratio of the peak heights for each residue in the NOE versus a reference spectrum determined the Heteronuclear {^{1}H}-^{15}N NOE. Uncertainties in the heteronuclear NOE were estimated by propagation of error using spectral noise using the formula:

$$\Delta NOE = [\{(1/I_{REF}) * \delta I_{NOE}\}^2 + \{(-I_{NOE} / I_{REF}^2) * \delta I_{REF}\}^2]^{½}$$

where ΔNOE is the propagation of error in the heteronuclear NOE, I_{NOE} and I_{REF} are the peak heights of the NOE and control reference spectrum, δI_{NOE} and δI_{REF} are spectral noise of the NOE and control reference spectrum, respectively. The residue-specific order parameter (S^2) was calculated using TENSOR2.[43]

Hydrogen exchange experiments

Amide proton-deuterium exchange rates were measured at 298 K and pH 6.5. The protein was lyophilized, dissolved into 100% D_2O[44] and a series of ^{15}N-^{1}H HSQC spectra were collected over a period of time to monitor the decay of the amide signals as the protium is exchanged by deuterium. The pseudo first-order rate constants for exchange, k_{ex}, were calculated using inhouse MATLAB codes by nonlinear least-squares fitting of the peak intensities, I_t (normalized by the number of transients) to the equation: $I_t = (I_0-I_{00}) * \exp(-k_{ex} * t) + I_{00}$, where t is the midpoint time of each spectrum, I_0 is the initial peak intensity and I_{00} accounts for the intensity coming from residual water.[45,46] Error in k_{ex} was determined by Monte Carlo simulation.

Amide proton–proton exchange rates were measured at pH 6.5 and 7.5 by the CLEANEX-PM method[47,48] at 303 K. At each pH, a series of spectra with 10, 20, 30, 40, 50, 60 and 80 ms transfer periods and a reference spectrum using a recycle delay of 12.0 sec were collected.

The pseudo first-order rate constants for chemical exchange, k_{ex}, were calculated by nonlinear least-squares fitting of the peak intensities versus transfer time using inhouse MATLAB codes. A scaling factor of 0.7 was used to correct for the steady-state water magnetization.

The protection factors (PFs) for each amide proton were calculated as the ratio of the predicted intrinsic exchange rate constant (k_{int}) for an unstructured polypeptide with the same amino acid sequence versus the experimentally calculated exchange rate constant (k_{ex}). The k_{int} values were calculated using the program Sphere.[49] Here an EX2 mechanism is assumed, where the exchange rate constants (k_{ex}) have a first-order dependence on sample pH and temperature. Also, the protein stability is assumed to be independent of pH and temperature.

Splicing activity assay of wildtype and mutant inteins

The 5' and 3' ends of *Spl* DnaX mini-intein gene were stitched by PCR with ubiquitin gene (N-extein) and a homeodomain gene (C-extein) of HOX proteins, respectively. This was cloned into pET28a(+) vector resulting in N-terminal His_6-tag. Additionally, a 3X-Flag tag was also added to the C-terminus of this construct. Several point mutations were introduced in the intein by inverse PCR based site-directed mutagenesis. These plasmids containing His_6-UBQ-Intein-HD-$Flag_3$ were transformed into *E. coli* BL21(λDE3) cells and grown in LB media supplemented with 70 µg/ml kanamycin at 37°C until the cell density reached OD_{600} of ~0.6-0.7. Protein expression was induced by 1 mM IPTG. Different time points were collected up to 1-hour post-induction. Collected cells were harvested by centrifugation and lysed with 1X SDS-PAGE loading dye containing 20% BME. The samples were boiled at 100°C for ~10 minutes, centrifuged and the supernatant from the lysate were run on a 12% SDS-PAGE.

For Western blot analysis, the gels were blotted onto PVDF (polyvinylidene difluoride) membrane with a pore size of 0.45 µm at 20 Volts for 20 mins. The blots were incubated for 1 hour in 5% non-fat dry milk powder solution in 1X TBST (tris-buffered saline with tween 20) at 25°C to remove any nonspecific binding and then with 1:5000 dilution of His_6-tag and Flag-tag mouse monoclonal antibody, separately, in 1X TBST for 16 hours at 4°C. The blots were washed with TBST three times at 10 mins interval, incubated with 1:10000 dilution of goat anti-mouse antibody for 1 hour at 25°C, and then washed again with TBST three times at 10 mins interval. The blots were developed for 5 min with chemiluminescent substrate Luminol for horseradish peroxidase (HRP). Comparison with the protein marker allowed identification of bands from each product. The band intensities were determined using the software ImageJ (US National Institutes of Health). The yields of the splicing and the cleavage products were normalized and summarized in Table 3. All the experiments were performed triplicates.

RESULTS

The *Spl* DnaX mini intein is catalytically active

To determine whether the *Spl* DnaX mini intein is catalytically active a His_6-UBQ-Intein-HD-$Flag_3$ protein construct (Figure 2A) consisting of *Spl* DnaX mini-intein (INT) with ubiquitin (UBQ) and homeodomain (HD) as the N-extein and C-extein, respectively, was designed and

expressed in *E. coli*. After induction, cells were collected at different time points, lysed and analyzed by SDS-PAGE followed by western blotting (Figure 2B). The 37.6 kDa precursor protein was not detected at any time, indicating rapid catalysis of the splicing reaction by the intein enzyme immediately upon synthesis in the cell. The ligated product (His_6-UBQ-HD-$Flag_3$, 21.9 kDa) along with the excised intein (INT, 15.7 kDa) were detected, confirming that the *Spl* DnaX intein is catalytically active. The precursor protein also undergoes side reactions where cleavage occurs at the N-terminus or the C-terminus of the intein (Figure 2A). C-terminal cleavage products, i.e. His_6-UBQ-INT (26.6 kDa) and HD-$Flag_3$ (11 kDa), were detected by western blotting (Figure 2B). One of the N-terminal cleavage product, i.e. His_6-UBQ (10.9 kDa) was also detected. Based on the relative band intensities of the spliced product and the side reaction products, the splicing efficiency of *Spl* DnaX mini-intein was ~60%.

The *Spl* DnaX mini intein is a monomer in solution

As the *Spl* DnaX mini-intein is highly active, the N-terminal His_6-tag gets cleaved within the cell, thus hampering affinity purification of the protein and resulting in poor yield of the purified protein. Hence, the residues C1 and C(+1) were mutated to alanines to enable high quality purification with good yield, which is necessary for structural studies.

Both $Intein^{WT}$ and $Intein^{C1A/C(+1)A}$ were obtained in high purity (Figure 3A) and were further characterized by analytical size exclusion chromatography (SEC) and analytical ultracentrifugation (AUC). The SEC profiles confirm high purity and monomeric state of both proteins (Figure 3B). The $Intein^{WT}$ and $Intein^{C1A/C(+1)A}$ elute at 12.49 ml and 12.37 ml, respectively. The $Intein^{C1A/C(+1)A}$ has 5 extra residues in the N-terminus compared to $Intein^{WT}$ (Table 1), which most likely results in the slightly different elution volumes.

From sedimentation velocity experiment in AUC, the Stokes radius of 2.14 nm and 2.21 nm, sedimentation coefficient (S_w) of 1.776 S and 1.819 S and frictional ratio (f/f_0) of 1.26 and 1.28 were determined for $Intein^{WT}$ and $Intein^{C1A/C(+1)A}$, respectively (Figure 3C and 3D). The S_w values correlates to molecular weights of 16.9 kDa and 17.7 kDa for $Intein^{WT}$ and $Intein^{C1A/C(+1)A}$, respectively, which closely matches with the expected molecular weights for the proteins. Additionally, the frictional ratio of ~1.3 strongly suggests that both are globular proteins.[50] The continuous c(s) distribution model for fitted data clearly indicates the presence of a monomeric protein as the dominant species in solution (Figure 3D). Overall, the SEC and AUC studies show that $Intein^{WT}$ and $Intein^{C1A/C(+1)A}$ have similar hydrodynamic properties, indicating no perturbation due to mutation.

Solution NMR structure of *Spl* DnaX mini intein

The structure of $Intein^{C1A/C(+1)A}$ in solution was determined at pH 6.5 by NMR spectroscopy. The ^{15}N-^{1}H HSQC spectrum of $Intein^{C1A/C(+1)A}$ (Figure 4A) showed well-dispersed peaks indicating the presence of a properly folded domain. The ^{15}N-^{1}H HSQC spectrum was assigned using standard $^{15}N/^{13}C/^{1}H$ heteronuclear NMR experiments. $Intein^{C1A/C(+1)A}$ contains 141 residues (Table 1). Nearly complete assignments of the backbone (99.85%) and the sidechain (98%) chemical shifts (except the aromatic atoms) were obtained. The

diastereotopic methyl groups present in leucines and valines were assigned by fractional (10%) ^{13}C-labeled sample.[51] Using the backbone $^{1}H^{N}$, ^{15}N, $^{13}C_{\alpha}$, $^{13}C_{\beta}$ and ^{13}CO chemical shifts, secondary structure propensity was calculated by the program MICS.[34] It shows the presence of thirteen beta strands along with two helices for Intein$^{C1A/C(+1)A}$ (Figure 4B). Intein$^{C1A/C(+1)A}$ has two cysteine residues C25 and C65 with $^{13}C_{\beta}$ chemical shifts of 31.07 and 34.05 ppm, respectively. This indicates that C25 and C65 do not form a disulfide linkage between them.[52] Similarly, based on C_{β} and C_{γ} chemical shifts,[53] all three proline residues in the protein were found to have *trans*-peptide bonds.

Distance and dihedral angle restraints were used to calculate the initial structures of the protein. Once consistent folded structures were obtained, hydrogen bond restraints, based on hydrogen exchange experiments, were also added. In the final step 100 structures were calculated, from which the statistics of the top 15 are summarized in Table 2. On average, 17.5 restraints were obtained for each residue. Backbone Φ and Ψ torsion angles of 97.1% residues are in favored regions of the Ramachandran plot.[54] None of the residues are in disallowed regions. These structures align with a root-mean-square deviation (RMSD) of 0.53 Å for all heavy atoms and 0.27 Å for backbone atoms across all ordered residues (Table 2). The regions with the highest degree of flexibility are located at the elongated loop between β10 and β11 strands (Figure 5). NMR assignment data has been deposited in the BioMagResBank (BMRB 50361), and the structural coordinates have been deposited in the Protein Data Bank (PDB 7CFV).

Amide hydrogen exchange reveals a highly stable core of *Spl* DnaX mini intein

Residue-wise stability of *Spl* DnaX mini intein was determined by NMR-based hydrogen exchange (HX) experiments. Using complementary protium-deuterium HX (pH 6.5) and protium-protium CLEANEX-PM (pH 6.5 and 7.5) experiments, the protection factors (PFs) for the backbone amide protons were calculated (Figure 6). HX measurements of Intein$^{C1A/C(+1)A}$ revealed a wide range of stability throughout the protein. Several amides did not display any significant decrease in peak intensity over a period of ~110 days, the course of the experiment. These amides have been assigned a PF of >10^8 (Figure 6, grey bars). Their high protection indicates that most likely these amides exchange via a global unfolding pathway. These amides are located mostly in the β-strands and form the rigid hydrophobic core of the enzyme. Residues in the 3_{10} helix (PF ~ 10^7), α-helix (PF ~ 10^3 to 10^6), β1 strand (PF ~ 10^6 to 10^7), and the turn between β1 and β2 (PF ~ 10^6 to 10^7) have intermediate stability indicating sub-global fluctuations. In contrast, amides in most loops have low stability (PF ~ 10 to 10^4) and most likely exchange via local conformational fluctuations. Particularly, the residues in the loop between β10 and β11 have very low stability (PF < 10). This is consistent with its solvent exposed structure (Figure 5C).

^{15}N relaxation measurements reveal fast dynamics of the *Spl* DnaX mini intein

Amide ^{15}N relaxation (R_1, R_2 and steady-state heteronuclear NOE) experiments were collected at 30C for Intein$^{C1A/C(+1)A}$ (Figure 7). For almost all residues, the heteronuclear {^{1}H}-^{15}N NOE

values are in the range of 0.55 to 0.88 (Figure 7A), indicating the presence of a well-folded structure.[42] The residues in the loop between β10 and β11 have low {^{1}H}-^{15}N NOE values, consistent with their low protection factors, and indicate local flexibility. The sidechain $N_{\varepsilon1}$-$H_{\varepsilon1}$ of all six tryptophans have {^{1}H}-^{15}N NOE values above 0.7, indicating that these sidechains have highly restricted mobility (Figure 7A). This is consistent with their buried conformation in the intein structure. Similarly, out of six arginine residues, the sidechain N_{ε}-H_{ε} of Arg17 and Arg73 have {^{1}H}-^{15}N NOE values of 0.75 and 0.6, respectively, indicating the rigid nature of these long arginine sidechains (Figure 7A). The Arg17 and Arg73 sidechains are within 4 Å distance of Asp6 and Glu33 sidechains, respectively. Salt-bridge interactions with neighbouring negatively charged residues explains the rigidity of these arginine sidechains. These interactions should also contribute to the stability of the intein structure.[55,56]

Transverse relaxation rate constants (R_2) have a value of ~15 s^{-1} for most of the residues (Figure 7B). Some residues have significantly higher R_2 values indicating the presence of slower microsecond-millisecond timescale motions. Most of these residues are present in the loop regions; Ile94 and Asp95 are present in the loop between β10 and β11 strands, His125 and Asn126 are present in the loop between β11 and β12 strands, and His69 is present in the conserved TXXH motif between β7 and β8 strands. Interestingly, the residues Thr53 and Trp54, which are present at the center of the β6-strand, also have high R_2 values. The longitudinal relaxation rate constants (R_1) has much less variation with an average value of ~1.4 s^{-1} (Figure 7C). The residue-specific order parameter (S^2), which gives a measure of the residue-wise rigidity of the protein, was calculated using TENSOR2.[43] Most of the residues, including the N and C termini, exhibit high order parameters. The average order parameter of the protein is 0.97 ± 0.03 (Figure 7D). Using the R_2/R_1 ratio, the isotropic tumbling correlation time, τ_c, for Intein$^{C1A/C(+1)A}$ was calculated to be 9.84 ± 0.03 ns. This is consistent with the molecular mass of the protein and confirms its monomeric state.

Identification of the active-site residues and their roles in catalysis

The three-dimensional structure of *Spl* DnaX mini-intein and previous work on other intein enzymes[19,57–61] suggested that the residues Cys1, Asn136 and Cys(+1) should be directly involved in the catalysis of the splicing reaction (Figure 1). To determine their catalytic roles, a series of mutations were designed where these three residues were mutated to alanines in all possible combinations of single, double and triple mutations (Table 1). The double and triple mutants were designed to study whether the catalytic functions of these active-site residues were coupled to each other. The effect of these mutations on the in cell splicing after one hour of induction was monitored by SDS-PAGE. The splicing as well as the N- and C-terminal cleavage products were specifically detected by Western blot using antibody against His_6-tag and are summarized in Table 3.

None of the mutants resulted in the spliced product, which indicates that these residues are indispensable for the protein splicing reaction catalyzed by *Spl* DnaX mini-intein (Figure 8). Substitution of Cys1 by Ala completely blocked the splicing and N-terminal cleavage reactions in all four C1A mutants. The C-terminal cleavage products, His_6-UBQ-INT (26.6 kDa) and HD-$Flag_3$ (11 kDa), were detected for UBQ-INT^{C1A}-HD and UBQ-INT$^{C1A/C(+1)A}$-HD. Thus, Cys1 is necessary for the initial nucleophilic attack to the preceding peptide bond and formation of the

linear thioester intermediate (Figure 1, step 1). The Cys(+1) to Ala mutation (UBQ-INT$^{C(+1)A}$-HD) prevents the splicing reaction but results in the cleavage at both N- and C-terminal of the intein enzyme. Thus, Cys(+1) is necessary for the formation of the branched thioester intermediate (Figure 1, step 2).

Mutation of Asn136 to Ala prevented the splicing and C-terminal cleavage reactions in all four N136A mutants. However, the N-terminal cleavage product His$_6$-UBQ was observed for UBQ-INT^{N136A}-HD and UBQ-INT$^{N136A/C(+1)A}$-HD (Figure 8). Thus, the C-terminal cleavage reaction is catalyzed by Asn136 (Figure 1, step 3). The triple mutant UBQ-INT$^{C1A/N136A/C(+1)A}$-HD yielded the stable precursor protein with molecular weight of 37.6 kDa which neither spliced, nor cleaved at the two termini. Thus, Cys1, Asn136 and Cys(+1) are the active-site residues, which are directly involved in the catalysis of the splicing reaction.

DISCUSSIONS

Spl DnaX mini-intein has a conserved HINT domain structure

The *Spl* DnaX mini-intein is one of the smallest intein enzymes (136 residues) characterized so far (Table S1). Its sequence was compared with other inteins, whose structure has been determined (Figure S1), and the highest sequence identity of 32% was obtained with *Ssp* DnaB (Table S1). Although sequence identity is low, *Spl* DnaX mini-intein has very good structural homology with other inteins (Table S1). The *Spl* DnaX mini-intein comprises of 13 β-strands, one α-helix, and one 3_{10}-helix, which are arranged in a compact horseshoe-like disk-shaped fold commonly found in HINT (hedgehog/intein) domain superfamily (Figure 5). The 3_{10}-helical turn was confirmed by observing a potential hydrogen bonding distance of ~2.4 Å in all the models between the carbonyl oxygen of Ala81 and amide proton of Ile84, thus, involving 10 backbone atoms in a turn. In inteins, the endonuclease domain is typically inserted within the HINT fold between β10 and β11 strands. In *Spl* DnaX mini-intein, the endonuclease domain is replaced by an extended loop between β10 and β11 strands. This loop does not affect the overall HINT fold of the *Spl* DnaX mini-intein. Thus, the structure of *Spl* DnaX mini-intein resembles very well with other intein structures.[62–67]

Based on the solved structure of *Spl* DnaX mini-intein and the sequence alignment with other inteins (Figure S1), we have identified four conserved blocks of residues i.e. block A, B, F and G, which are characteristic of all inteins (Perler 2002).[17] Block A begins with the conserved Cys nucleophile (C1) and consists of 13 residues. Usually 60-100 residues from the N-terminus, Block B is present and contains a conserved TXXH motif. Residues T66 and H69 form this motif in *Spl* DnaX mini-intein. Blocks C, D, E and H are usually present in the homing endonuclease domain and hence, not observed in *Spl* DnaX mini-intein. Block F contains a conserved Asp and a conserved HNF motif which are D118 and H125-N126-F127 in *Spl* DnaX mini-intein. Lastly, Block G ends with a conserved HNC motif, which are residues H135-N136-C(+1) in *Spl* DnaX mini-intein.

In *Spl* DnaX mini-intein the β-strands β6(43-59) and β11(100-122) form a long, curving antiparallel β-sheet, typical of intein enzymes. Extreme curvature of these strands puts high strain on this structure, which is most likely relieved by dynamic motions. In the strand β6, the backbone amides of the residues Thr53 and Trp54 exhibit motions in the slow millisecond-microsecond timescale, characteristic of conformational exchange.[68] On the other hand, Gly112 in the β11 strand exhibits faster motion in the picosecond-nanosecond timescale resulting in an order parameter of 0.62 (Figure 7D). Similarly, from the hydrogen exchange experiments, the several residues in the β6 and β11 strands have relatively less protection factors from the rest of the protein (Figure 6). Similar dynamic motion and low protection factors in the corresponding β-strands have also been observed for *Mtu* RecA[69] and *Pab* PolII[70] inteins. Moreover, few loop residues in *Spl* DnaX mini-intein such as His69, present in the conserved TXXH motif, Ile94 and Asp95, present in the loop between β10 and β11, His125 and Asn126, present in the loop between β11 and β12, have significantly higher R_2 relaxation rate constants (Figure 7B). This indicates slow microsecond-millisecond timescale motion resulting from conformational exchange. Similar dynamic motions in the corresponding regions have been observed for *Mja* KlbA,[71] *Npu* DnaE,[64] *Pho* RadA [72] and *Mtu* RecA[69] inteins. Thus, these dynamic motions appear to be a conserved feature of the intein structure.

Mechanism of splicing reaction mediated by *Spl* DnaX mini-intein

Based on their reaction mechanisms (Figure 1), intein enzymes have been categorized into three classes.[18] The *Spl* DnaX mini-intein has four cysteines Cys1, Cys25, Cys65 and Cys(+1). Mutation of Cys1 and Cys(+1), alone or in combination, completely abolished the splicing reaction (Figure 8). Thus, the two terminal cysteines, and not the intermediary Cys25 and Cys65, are involved in catalyzing the splicing reaction, confirming that *Spl* DnaX mini-intein belongs to Type I class of inteins.[18] Moreover, unlike the class 3 inteins which have a conserved Cys present in the Block F position,[18] the presence of the intermediary Cys65 in the Block B and no characteristic Cys in the Block F of *Spl* DnaX mini-intein (Figure S1) further confirms its class 1 nature. The Cys1 to Ala mutation in all four mutants prevented the nucleophilic attack of the thiol group of Cys1 to the carbonyl group of the previous peptide bond between Glu(-1) and Cys1 and thus, blocked the N-S acyl rearrangement, which prevents the formation of the linear thioester intermediate (Figure 1A, step 1). Attack on the linear thioester intermediate by the Cys(+1) thiol results in the formation of the branched thioester intermediate and subsequently the spliced product (Figure 1A, step 2). Hydrolysis of the thioester intermediates results in N-terminal cleavage. Hence, Cys1 to Ala mutation prevented both the splicing and N-terminal cleavage reactions (Table 3).

The Cys(+1) thiol attacks the linear thioester intermediate resulting in the formation of the branched intermediate (Figure 1, step 2). Both UBQ-INT^{N136A}-HD and UBQ-INT$^{N136A/C(+1)A}$-HD mutants are incapable of splicing and C-terminal cleavage reactions due to the absence of Asn136. However, UBQ-INT^{N136A}-HD can form both linear and branched intermediates whereas UBQ-INT$^{N136A/C(+1)A}$-HD can form only the linear intermediate. Hydrolysis of these intermediates results in the N-terminal cleavage products. As UBQ-INT^{N136A}-HD can form both intermediates, it results in a higher amount of N-terminal cleavage product relative to UBQ-INT$^{N136A/C(+1)A}$-HD (Table 3). Consistent with this, the UBQ-INT$^{C(+1)A}$-HD mutant undergoes relatively less N-terminal cleavage compared to UBQ-INT^{N136A}-HD (Table 3).

The Asn136 to Ala mutation prevents the splicing and C-terminal cleavage reactions, indicating that Asn136 is necessary for C-terminal cleavage step (Figure 1A, step 3). Interestingly, the C-terminal cleavage product is consistently low in the UBQ-INT$^{C(+1)A}$-HD and UBQ-INT$^{C1A/C(+1)A}$-HD mutants compared to UBQ-INT^{C1A}-HD mutant (Table 3). This indicates that Cys(+1) most likely facilitates the Asn cyclization and subsequent C-terminal cleavage (Figure 1A, step 3). Collectively, these mutagenesis experiments define the roles of the three conserved splice junction residues.

Furthermore, alignment of the *Spl* DnaX intein structure (Table 3) and sequence (Figure S1) with other inteins helped in the identification of more residues that may facilitate the catalysis of the splicing reaction. Thr66, which belongs to the TXXH motif in block B, is in close contact with the N-terminal scissile peptide bond and most likely helps in the N-S acyl rearrangement reaction.[73] His69, also part of the TXXH motif, facilitates the nucleophilic attack by C1, possibly by distorting the preceding peptide bond.[74] Asp118 facilitates the nucleophilic attack by Cys1 by lowering the activation energy and stabilizing the negative charge of Cys1 sidechain.[75] Asp118 also helps in the trans-esterification reaction by accepting a proton from the Cys(+1) sidechain.[76] His125 activates the Asn136 sidechain for nucleophilic attack.[62] His135 and Asp118 form part of an oxyanion hole and stabilize the tetrahedral intermediate, necessary for C-terminal asparagine cyclization.[62]

Competition between splicing and cleavage reactions in *Spl* DnaX mini-intein

Apart from the splicing reaction, inteins also undergo N- and C-terminal cleavage reactions. Since inteins are single turnover enzymes, these reactions decrease the overall yield of the spliced product. We designed all possible alanine mutations of the active site residues C1, N136 and C(+1) of *Spl* DnaX intein and functionally characterized them. These studies also provided important insights into the competition between the splicing and the cleavage reactions and the underlying roles of the active site residues.

Splicing occurred only in the wildtype enzyme and requires all three active site residues. Among the cleavage reactions, the C-terminal cleavage is highly efficient as evident from the single mutants (Table 3). For the N-terminal cleavage incompetent mutant UBQ-INT^{C1A}-HD, 90% C-terminal cleavage was detected, whereas for the C-terminal cleavage incompetent mutant UBQ-INT^{N136A}-HD only 7.5% N-terminal cleavage was detected. Thus, it appears that the N- and C-terminal cleavage reactions can proceed independently in *Spl* DnaX intein mutants. This is consistent with similar observations for *Ssp* DnaB intein.[62] It is important to note that according to the class 1 reaction mechanism (Figure 1), the C-terminal cleavage should occur after the N-terminal nucleophilic attack and the formation of the branched intermediate in order to form the spliced product. A precursor protein undergoing C-terminal cleavage before branched intermediate formation will not produce the spliced product. Hence, the suppression of premature C-terminal cleavage is necessary for a successful splicing reaction. In the wildtype *Spl* DnaX intein, the spliced product is indeed the major product (~60%). Interestingly, the N- and C-terminal cleavage products are almost equal (~20%) for the wildtype intein. This indicates that either cysteines C1 and C(+1) suppress the reactivity of N136 or N136 enhances the reactivity of the cysteines in the wild type intein and thus, favour

the formation of the spliced product. Overall, this highlights the interdependencies of the catalytic residues in driving the splicing reaction.

CONCLUSIONS

In summary, we have determined the solution NMR structure of a new mini-intein enzyme from the cyanobacterium *Spirulina platensis*. We show that the *Spl* DnaX mini-intein is monomeric in solution and has a typical HINT domain fold. We also characterized the dynamics of the protein and found conserved motions in the protein. NMR-based hydrogen exchange experiments revealed the presence of a highly stable core in the *Spl* DnaX mini-intein. In cell splicing assays demonstrated that *Spl* DnaX mini-intein is a highly active enzyme. We further investigated the functional roles of the catalytic residues by designing combinatorial alanine mutations, which showed their combined effects in suppressing N- and C-terminal cleavage reactions and enhancing the splicing product. Thus, we report the detailed characterization of a new intein enzyme, which can be used for various intein-based applications.

ACKNOWLEDGEMENTS

This work was supported by the Science and Engineering Research Board (SERB), India Early Career Research (ECR) award (ECR/2016/000847 dated 07/03/2017), and ISIRD grant from IIT-Kharagpur (IIT/SRIC/ISIRD/2015-2016 dated 14/12/2015) to S.D. NSF award DBI-1902076, the CU-Denver STYPE-61193205 and NIH grant P41GM103399 (NIGMS) supported W.L. The authors would like to thank the Central Research Facility (CRF) at IIT Kharagpur for the use of NMR, SEC and AUC facilities.

CONFLICT OF INTEREST

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

REFERENCES

1 R. Hirata, Y. Ohsumk, A. Nakano, H. Kawasaki, K. Suzuki and Y. Anraku, Molecular structure of a gene, *VMA1*, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*, *J. Biol. Chem.*, 1990, **265**, 6726–6733.

2 P. M. Kane, C. T. Yamashiro, D. F. Wolczyk, N. Neff, M. Goebl and T. H. Stevens, Protein splicing converts the yeast *TFP1* gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase, *Science*, 1990, **250**, 651–657.

3 M. Q. Xu, M. W. Southworth, F. B. Mersha, L. J. Hornstra and F. B. Perler, In vitro protein splicing of purified precursor and the identification of a branched intermediate, *Cell*, 1993, **75**, 1371–1377.

4 H. Paulus, Protein splicing and related forms of protein autoprocessing, *Annu. Rev. Biochem.*, 2000, **69**, 447–496.

5 X. Q. Liu, Protein-splicing intein: Genetic mobility, origin, and evolution, *Annu. Rev. Genet.*, 2000, **34**, 61–76.

6 G. Volkmann and H. D. Mootz, Recent progress in intein research: from mechanism to directed evolution and applications, *Cell. Mol. Life Sci.*, 2013, **70**, 1185–1206.

7 D. W. Wood and J. A. Camarero, Intein applications: from protein purification and labeling to metabolic control methods, *J. Biol. Chem.*, 2014, **289**, 14512–14519.

8 N. H. Shah and T. W. Muir, Inteins: Nature's gift to protein chemists, *Chem. Sci.*, 2014, **5**, 446–461.

9 T. W. Muir, D. Sondhi and P. A. Cole, Expressed protein ligation: A general method for protein engineering, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 6705–6710.

10 T. W. Muir, Semisynthesis of proteins by expressed protein ligation., *Annu. Rev. Biochem.*, 2003, **72**, 249–289.

11 H. D. Mootz, E. S. Blum, A. B. Tyszkiewicz and T. W. Muir, Conditional Protein Splicing: A New Tool to Control Protein Structure and Function in Vitro and in Vivo, *J. Am. Chem. Soc.*, 2003, **125**, 10561–10569.

12 H. D. Mootz, E. S. Blum and T. W. Muir, Activation of an Autoregulated Protein Kinase by Conditional Protein Splicing, *Angew. Chemie Int. Ed.*, 2004, **43**, 5189–5192.

13 T. Kurpiers and H. D. Mootz, Site-Specific Chemical Modification of Proteins with a Prelabelled Cysteine Tag Using the Artificially Split Mxe GyrA Intein, *ChemBioChem*, 2008, **9**, 2317–2325.

14 D. Cowburn and T. W. Muir, Segmental Isotopic Labeling Using Expressed Protein Ligation, *Nucl. Magn. Reson. Biol. Macromol. - Part B*, 2001, **339**, 41–54.

15 D. Liu, R. Xu and D. Cowburn, in *Methods in Enzymology*, Elsevier Inc., 1st edn., 2009, vol. 462, pp. 151–175.

16 T. C. Evans, J. Benner and M. Xu, The Cyclization and Polymerization of Bacterially Expressed Proteins Using Modified Self-splicing Inteins, *J. Biol. Chem.*, 1999, **274**, 18359–18363.

17 F. B. Perler, InBase: the Intein Database, *Nucleic Acids Res.*, 2002, **30**, 383–384.

18 K. Tori, B. Dassa, M. A. Johnson, M. W. Southworth, L. E. Brace, Y. Ishino, S. Pietrokovski and F. B. Perler, Splicing of the Mycobacteriophage Bethlehem DnaB intein: identification of a new mechanistic class of inteins that contain an obligate block F nucleophile, *J. Biol. Chem.*, 2010, **285**, 2515–2526.

19 M. Q. Xu and F. B. Perler, The mechanism of protein splicing and its modulation by mutation, *EMBO J.*, 1996, **15**, 5146–5153.

20 M. W. Southworth, An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile, *EMBO J.*, 2000, **19**, 5019–5026.

21 L. Saleh, M. W. Southworth, N. Considine, C. O'Neill, J. Benner, J. M. Bollinger and F. B. Perler, Branched Intermediate Formation Is the Slowest Step in the Protein Splicing Reaction of the Ala1 KlbA Intein from *Methanococcus jannaschii*, *Biochemistry*, 2011, **50**, 10576–10589.

22 L. E. Brace, M. W. Southworth, K. Tori, M. L. Cushing and F. Perler, The *Deinococcus radiodurans* Snf2 intein caught in the act: Detection of the Class 3 intein signature Block F branched intermediate, *Protein Sci.*, 2010, **19**, 1525–1533.

23 X. Duan, F. S. Gimble and F. A. Quiocho, Crystal Structure of PI-SceI, a Homing Endonuclease with Protein Splicing Activity, *Cell*, 1997, **89**, 555–564.

24 T. Klabunde, S. Sharma, A. Telenti, W. R. Jacobs and J. C. Sacchettini, Crystal structure of GyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing, *Nat. Struct. Biol.*, 1998, **5**, 31–36.

25 B. W. Poland, M. Xu and F. A. Quiocho, Structural Insights into the Protein Splicing Mechanism of PI-SceI, *J. Biol. Chem.*, 2000, **275**, 16408–16413.

26 R. Mizutani, S. Nogami, M. Kawasaki, Y. Ohya, Y. Anraku and Y. Satow, Protein-splicing Reaction via a Thiazolidine Intermediate: Crystal Structure of the *VMA1*-derived Endonuclease Bearing the N and C-terminal Propeptides, *J. Mol. Biol.*, 2002, **316**, 919–929.

27 R. J. Goldberg, Sedimentation in the Ultracentrifuge, *J. Phys. Chem.*, 1953, **57**, 194–202.

28 P. Schuck, Size-Distribution Analysis of Macromolecules by Sedimentation Velocity Ultracentrifugation and Lamm Equation Modeling, *Biophys. J.*, 2000, **78**, 1606–1619.

29 C. A. Brautigam, in *Analytical Ultracentrifugation*, Elsevier Inc., 1st edn., 2015, vol. 562, pp. 109–133.

30 W. Lee, M. Tonelli and J. L. Markley, NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy, *Bioinformatics*, 2015, **31**, 1325–1327.

31 M. Sattler, Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients, *Prog. Nucl. Magn. Reson. Spectrosc.*, 1999, **34**, 93–158.

32 F. Delaglio, S. Grzesiek, G. Vuister, G. Zhu, J. Pfeifer and A. Bax, NMRPipe: A multidimensional spectral processing system based on UNIX pipes, *J. Biomol. NMR*, 1995, **6**, 277–293.

33 W. Lee, W. M. Westler, A. Bahrami, H. R. Eghbalnia and J. I. Markley, PINE-SPARKY: Graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy, *Bioinformatics*, 2009, **25**, 2085–2087.

34 Y. Shen and A. Bax, Identification of helix capping and β-turn motifs from NMR chemical shifts, *J. Biomol. NMR*, 2012, **52**, 211–232.

35 W. Lee, J. H. Kim, W. M. Westler and J. L. Markley, PONDEROSA, an automated 3D-NOESY peak picking program, enables automated protein structure determination, *Bioinformatics*, 2011, **27**, 1727–1728.

36 C. D. Schwieters, J. J. Kuszewski, N. Tjandra and G. Marius Clore, The Xplor-NIH NMR molecular structure determination package, *J. Magn. Reson.*, 2003, **160**, 65–73.

37 W. Lee, C. M. Petit, G. Cornilescu, J. L. Stark and J. L. Markley, The AUDANA algorithm for automated protein 3D structure determination from NMR NOE data, *J. Biomol. NMR*, 2016, **65**, 51–57.

38 W. Lee, J. L. Stark and J. L. Markley, PONDEROSA-C/S: client–server based software package for automated protein 3D structure determination, *J. Biomol. NMR*, 2014, **60**, 73–75.

39 W. Lee, G. Cornilescu, H. Dashti, H. R. Eghbalnia, M. Tonelli, W. M. Westler, S. E. Butcher, K. A. Henzler-Wildman and J. L. Markley, Integrative NMR for biomolecular research, *J. Biomol. NMR*, 2016, **64**, 307–332.

40 W. Kabsch and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 1983, **22**, 2577–2637.

41 S. De, A. C. K. Chan, H. J. Coyne, N. Bhachech, U. Hermsdorf, M. Okon, M. E. P. Murphy, B. J. Graves and L. P. McIntosh, Steric Mechanism of Auto-Inhibitory Regulation of Specific and Non-Specific DNA Binding by the ETS Transcriptional Repressor ETV6, *J. Mol. Biol.*, 2014, **426**, 1390–1406.

42 S. Maiti, B. Acharya, V. S. Boorla, B. Manna, A. Ghosh and S. De, Dynamic Studies on Intrinsically Disordered Regions of Two Paralogous Transcription Factors Reveal Rigid Segments with Important Biological Functions, *J. Mol. Biol.*, 2019, **431**, 1353–1369.

43 P. Dosset, J. C. Hus, M. Blackledge and D. Marion, Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data, *J. Biomol. NMR*, 2000, **16**, 23–28.

44 G. P. Connelly and L. P. McIntosh, Characterization of a Buried Neutral Histidine in *Bacillus circulans* Xylanase: Internal Dynamics and Interaction with a Bound Water Molecule, *Biochemistry*, 1998, **37**, 1810–1818.

45 H. J. Coyne, S. De, M. Okon, S. M. Green, N. Bhachech, B. J. Graves and L. P. McIntosh, Autoinhibition of ETV6 (TEL) DNA Binding: Appended Helices Sterically Block the ETS Domain, *J. Mol. Biol.*, 2012, **421**, 67–84.

46 S. De, M. Okon, B. J. Graves and L. P. McIntosh, Autoinhibition of ETV6 DNA Binding Is Established by the Stability of Its Inhibitory Helix, *J. Mol. Biol.*, 2016, **428**, 1515–1530.

47 T. L. Hwang, P. C. M. Van Zijl and S. Mori, Accurate quantitation of water-amide proton exchange rates using the Phase-Modulated CLEAN chemical EXchange (CLEANEX-PM) approach with a Fast-HSQC (FHSQC) detection scheme, *J. Biomol. NMR*, 1998, **11**, 221–226.

48 T. Hwang, S. Mori, A. J. Shaka and P. C. M. van Zijl, Application of Phase-Modulated CLEAN Chemical EXchange Spectroscopy (CLEANEX-PM) to Detect Water−Protein Proton Exchange and Intermolecular NOEs, *J. Am. Chem. Soc.*, 1997, **119**, 6203–6204.

49 Y.-Z. Zhang, Protein and peptide structure and interactions studied by hydrogen exchange and NMR, PhD Thesis, University of Pennsylvania, PA, USA, 1995.

50 H. P. Erickson, Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy, *Biol. Proced. Online*, 2009, **11**, 32–51.

51 D. Neri, T. Szyperski, G. Otting, H. Senn and K. Wüthrich, Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ^{13}C labeling, *Biochemistry*, 1989, **28**, 7510–7516.

52 D. Sharma and K. Rajarathnam, ^{13}C NMR chemical shifts can predict disulfide bond formation, *J. Biomol. NMR*, 2000, **18**, 165–171.

53 M. Schubert, D. Labudde, H. Oschkinat and P. Schmieder, A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on ^{13}C chemical shift statistics, *J. Biomol. NMR*, 2002, **24**, 149–154.

54 G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, Stereochemistry of polypeptide chain configurations, *J. Mol. Biol.*, 1963, **7**, 95–99.

55 D. J. Barlow and J. M. Thornton, Ion-pairs in proteins, *J. Mol. Biol.*, 1983, **168**, 867–885.

56 J. E. Donald, D. W. Kulp and W. F. DeGrado, Salt bridges: Geometrically specific, designable interactions, *Proteins Struct. Funct. Bioinforma.*, 2011, **79**, 898–915.

57 H. Paulus, Inteins as Enzymes, *Bioorg. Chem.*, 2001, **29**, 119–129.

58 T. C. Evans and M. Xu, Mechanistic and Kinetic Considerations of Protein Splicing, *Chem. Rev.*, 2002, **102**, 4869–4884.

59 L. Saleh and F. B. Perler, Protein splicing in cis and in trans, *Chem. Rec.*, 2006, **6**, 183–193.

60 K. V. Mills and H. Paulus, in *Homing Endonucleases and Inteins*, ed. Marlene Belfort et al, Springer-Verlag, Berlin/Heidelberg, 2005, pp. 233–255.

61 F. Perler, Protein Splicing Mechanisms and Applications, *IUBMB Life (International Union Biochem. Mol. Biol. Life)*, 2005, **57**, 469–476.

62 Y. Ding, M. Xu, I. Ghosh, X. Chen, S. Ferrandon, G. Lesage and Z. Rao, Crystal Structure of a Mini-intein Reveals a Conserved Catalytic Module Involved in Side Chain Cyclization of Asparagine during Protein Splicing, *J. Biol. Chem.*, 2003, **278**, 39133–39142.

63 A. S. Aranko, J. S. Oeemig, D. Zhou, T. Kajander, A. Wlodawer and H. Iwaï, Structure-based engineering and comparison of novel split inteins for protein ligation, *Mol. BioSyst.*, 2014, **10**, 1023–1034.

64 J. S. Oeemig, A. S. Aranko, J. Djupsjöbacka, K. Heinämäki and H. Iwaï, Solution structure of DnaE intein from *Nostoc punctiforme*: Structural basis for the design of a new split intein suitable for site-specific chemical modification, *FEBS Lett.*, 2009, **583**, 1451–1456.

65 P. Sun, S. Ye, S. Ferrandon, T. C. Evans, M. Xu and Z. Rao, Crystal Structures of an Intein from the Split *dnaE* Gene of *Synechocystis* sp. PCC6803 Reveal the Catalytic Model Without the Penultimate Histidine and the Mechanism of Zinc Ion Inhibition of Protein Splicing, *J. Mol. Biol.*, 2005, **353**, 1093–1105.

66 H. M. Beyer, K. M. Mikula, M. Li, A. Wlodawer and H. Iwaï, The crystal structure of the naturally split gp41-1 intein guides the engineering of orthogonal split inteins from *cis*-splicing inteins, *FEBS J.*, 2020, **287**, 1886–1898.

67 P. Van Roey, B. Pereira, Z. Li, K. Hiraga, M. Belfort and V. Derbyshire, Crystallographic and Mutational Studies of *Mycobacterium tuberculosis* recA Mini-inteins Suggest a Pivotal Role for a Highly Conserved Aspartate Residue, *J. Mol. Biol.*, 2007, **367**, 162–173.

68 A. G. Palmer, C. D. Kroenke and J. Patrick Loria, in *Methods Enzymol.*, 2001, vol. 339, pp. 204–238.

69 Z. Du, Y. Liu, D. Ban, M. M. Lopez, M. Belfort and C. Wang, Backbone Dynamics and Global Effects of an Activating Mutation in Minimized *Mtu* RecA Inteins, *J. Mol. Biol.*, 2010, **400**, 755–767.

70 Z. Du, J. Liu, C. D. Albracht, A. Hsu, W. Chen, M. D. Marieni, K. M. Colelli, J. E. Williams, J. N. Reitter, K. V Mills and C. Wang, Structural and Mutational Studies of a Hyperthermophilic Intein from DNA Polymerase II of *Pyrococcus abyssi*, *J. Biol. Chem.*, 2011, **286**, 38638–38648.

71 M. A. Johnson, M. W. Southworth, T. Herrmann, L. Brace, F. B. Perler and K. Wüthrich, NMR structure of a KlbA intein precursor from Methanococcus jannaschii, *Protein Sci.*, 2007, **16**, 1316–1328.

72 J. S. Oeemig, D. Zhou, T. Kajander, A. Wlodawer and H. Iwaï, NMR and Crystal Structures of the *Pyrococcus horikoshii* RadA Intein Guide a Strategy for Engineering a Highly Efficient and Promiscuous Intein, *J. Mol. Biol.*, 2012, **421**, 85–99.

73 E. Eryilmaz, N. H. Shah, T. W. Muir and D. Cowburn, Structural and Dynamical Features of

Inteins and Implications on Protein Splicing, *J. Biol. Chem.*, 2014, **289**, 14506–14511.
74 J. Binschik and H. D. Mootz, Chemical Bypass of Intein-Catalyzed N-S Acyl Shift in Protein Splicing, *Angew. Chemie Int. Ed.*, 2013, **52**, 4260–4264.
75 Z. Du, Y. Zheng, M. Patterson, Y. Liu and C. Wang, pK(a) Coupling at the Intein Active Site: Implications for the Coordination Mechanism of Protein Splicing with a Conserved Aspartate, *J. Am. Chem. Soc.*, 2011, **133**, 10275–10282.
76 B. Pereira, P. T. Shemella, G. Amitai, G. Belfort, S. K. Nayak and M. Belfort, Spontaneous Proton Transfer to a Conserved Intein Residue Determines On-Pathway Protein Splicing, *J. Mol. Biol.*, 2011, **406**, 430–442.
77 A. Bhattacharya, R. Tejero and G. T. Montelione, Evaluating protein structures determined by structural genomics consortia, *Proteins Struct. Funct. Bioinforma.*, 2006, **66**, 778–795.

TABLES

Table 1 List of *Spl* DnaX mini-intein constructs

Name of construct	Description
Intein^{WT}	Residues before Cys1 and after Asn136 are spontaneously cleaved off due to its enzymatic action. Total residues: 136
Intein$^{C1A/C(+1)A}$	Both Cys1 and Cys(+1) are mutated to Ala. The residues GSHM prior to Glu(-1) are left after thrombin cleavage. Total residues: 141
UBQ-INT^{WT}-HD	Fusion of N- and C-terminus of *Spl* DnaX intein protein with Ubiquitin (N-extein) and HD (C-extein) protein, respectively. The N-terminus has His_6 tag and C-terminus has 3X-Flag tag.
UBQ-INT^{C1A}-HD	Contains mutation of Cys1 to alanine in UBQ-INT-HD construct.
UBQ-INT^{N136A}-HD	Contains mutation of Asn136 to alanine in UBQ-INT-HD construct.
UBQ-INT$^{C(+1)A}$-HD	Contains mutation of Cys(+1) to alanine in UBQ-INT-HD construct.
UBQ-INT$^{C1A/N136A}$-HD	Contains mutations of Cys1 and Asn136 to alanines in UBQ-INT-HD construct.
UBQ-INT$^{C1A/C(+1)A}$-HD	Contains mutations of Cys1 and Cys(+1) to alanines in UBQ-INT-HD construct.
UBQ-INT$^{N136A/C(+1)A}$-HD	Contains mutations of Asn136 and Cys(+1) to alanines in UBQ-INT-HD construct.
UBQ-INT$^{C1A/N136A/C(+1)A}$-HD	Contains mutations of Cys1, Asn136 and Cys(+1) to alanines in UBQ-INT-HD construct.

Table 2 Experimental data and statistics for the solution NMR structure calculation of *Spl* DnaX mini-intein ensembles

Parameters	**Value**
(a) *NMR distance and dihedral restraints*	
Distance restraints	
Total number of NOE	2108
Short-range (|i-j| <= 1)	984
Medium-range (1 < |i-j| < 5)	220
Long-range (|i-j| >= 5)	904
Hydrogen bond restraints	56
Dihedral angle restraints	
Total	238
Phi (φ)	117
Psi (ψ)	121
Total number of restricting constraints	2402
(b) *Average RMSD (Å) against the lowest-energy conformer for ordered residues*[a]	
Backbone atoms (N, C^{α}, C, O)	0.277
All heavy atoms	0.528
(c) *RMSD from ideal geometry*	
For bond lengths (Å)	0.015
For bond angles (°)	1.4
(d) *Violations*	
Distance constraints (>0.5 Å)	0
Dihedral angle constraints (>5°)	0
Van der Waals constraints (>0.2 Å)	0
(e) *Ramachandran plot statistics (%) from PROCHECK for selected residues*[b]	
Residues in most favoured regions	90.3
Residues in additionally allowed regions	9.5
Residues in generously allowed regions	0.2
Residues in disallowed regions	0

(f) *Ramachandran Plot statistics (%) from MolProbity for selected residues*[b]		
Residues in most favourable regions	97.1	
Residues in additionally allowed regions	2.9	
Residues in disallowed regions	0	
(g) *Structure quality factors*	Raw score	Z-score[c]
Verify3D	0.16	-4.82
ProsaII	0.42	-0.95
MolProbity Clashscore	26.67	-3.05

[a]Ordered CYRANGE residues: 1-137.
[b]Selected residues: 2-66, 69-136.
[c]With respect to mean and SD for a set of 252 X-ray structures < 500 residues, of resolution <= 1.80 Å, R-factor <= 0.25 and R-free <= 0.28; a positive value indicates a 'better' score.
Structure quality was evaluated using PSVS 1.5.[77]

Table 3 Relative quantification of the splicing and cleavage products of the wildtype and mutant *Spl* DnaX intein

Name of Construct	N-terminal Cleavage	C-terminal Cleavage	Spliced Product	Precursor protein
UBQ-INT^{WT}-HD	21.6%	19.8%	58.6%	ND
UBQ-INT^{C1A}-HD	ND	90.5%	ND	9.5%
UBQ-INT^{N136A}-HD	7.5%	ND	ND	92.5%
UBQ-INT$^{C(+1)A}$-HD	0.5%	67.8%	ND	31.7%
UBQ-INT$^{C1A/N136A}$-HD	ND	ND	ND	100%
UBQ-INT$^{C1A/C(+1)A}$-HD	ND	70.2%	ND	29.8%
UBQ-INT$^{N136A/C(+1)A}$-HD	0.2%	ND	ND	99.8%
UBQ-INT$^{C1A/N136A/C(+1)A}$-HD	ND	ND	ND	100%

ND is not detected.
Quantification of the bands is based on anti-His_6 western blot images at 60 minutes after induction.

FIGURES

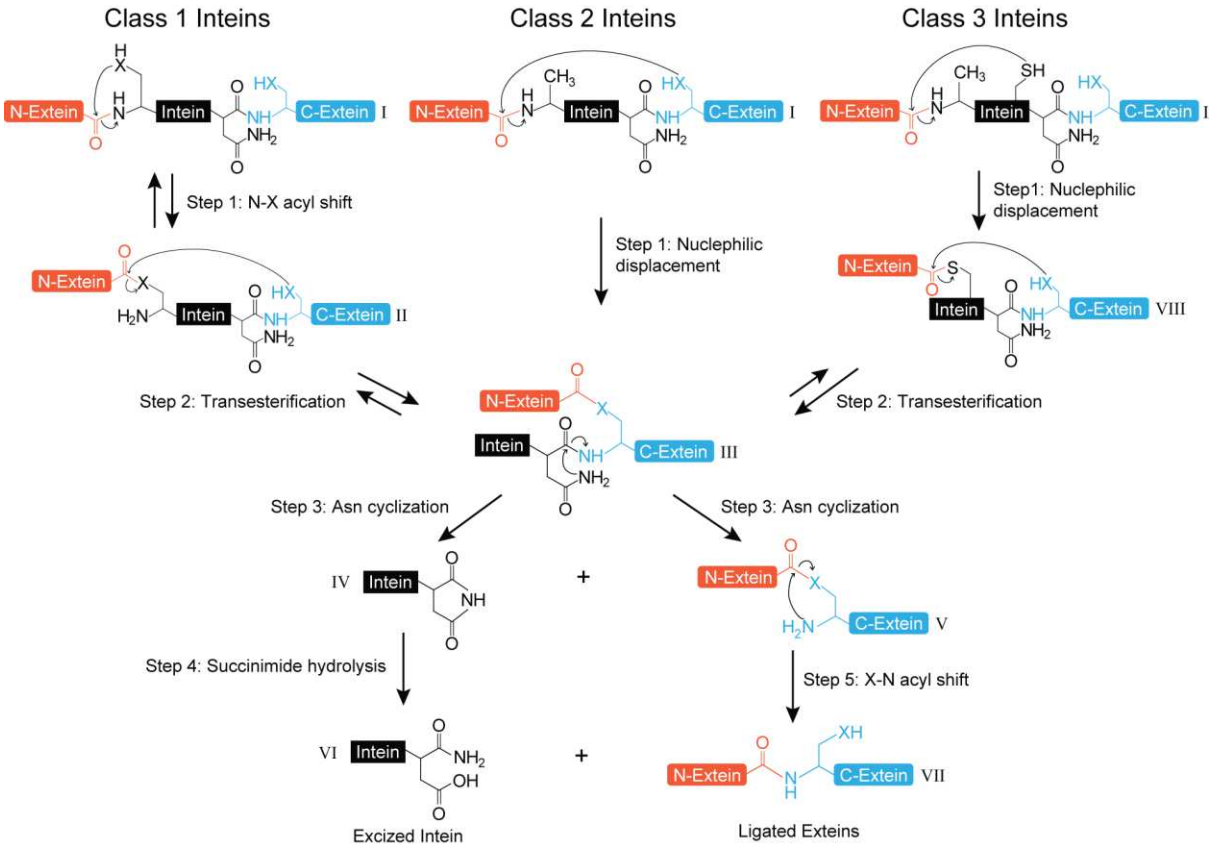


Figure 1 Catalytic mechanisms of splicing reactions by three different classes of inteins

Class 1 protein splicing pathway consists of four nucleophilic rearrangement reactions resulting in the spliced product (VII). The class 2 and class 3 inteins lack the N-terminal nucleophile. The first C-extein residue of class 2 inteins directly attacks the peptide bond at the N-terminal splice junction, resulting in the formation of the standard branched intermediate (III). In class 3 inteins, the sidechain nucleophile of Cys in the WCT triplet attacks the N-terminal peptide bond to form a branched intermediate (VIII). The N-extein is then transferred to the sidechain of the first C-extein residue resulting in the formation of the same branched intermediate (III) as in class 1 and 2 inteins. Once the standard branched intermediate is formed, the remaining steps are the same in all three classes of inteins. The precursor protein, linear ester intermediate, cyclized intein and ester-linked exteins are shown as I, II, IV and V, respectively. 'X' represents the sulfur or oxygen atom in the sidechain of Cys, Ser or Thr. Tetrahedral intermediates are not shown here.

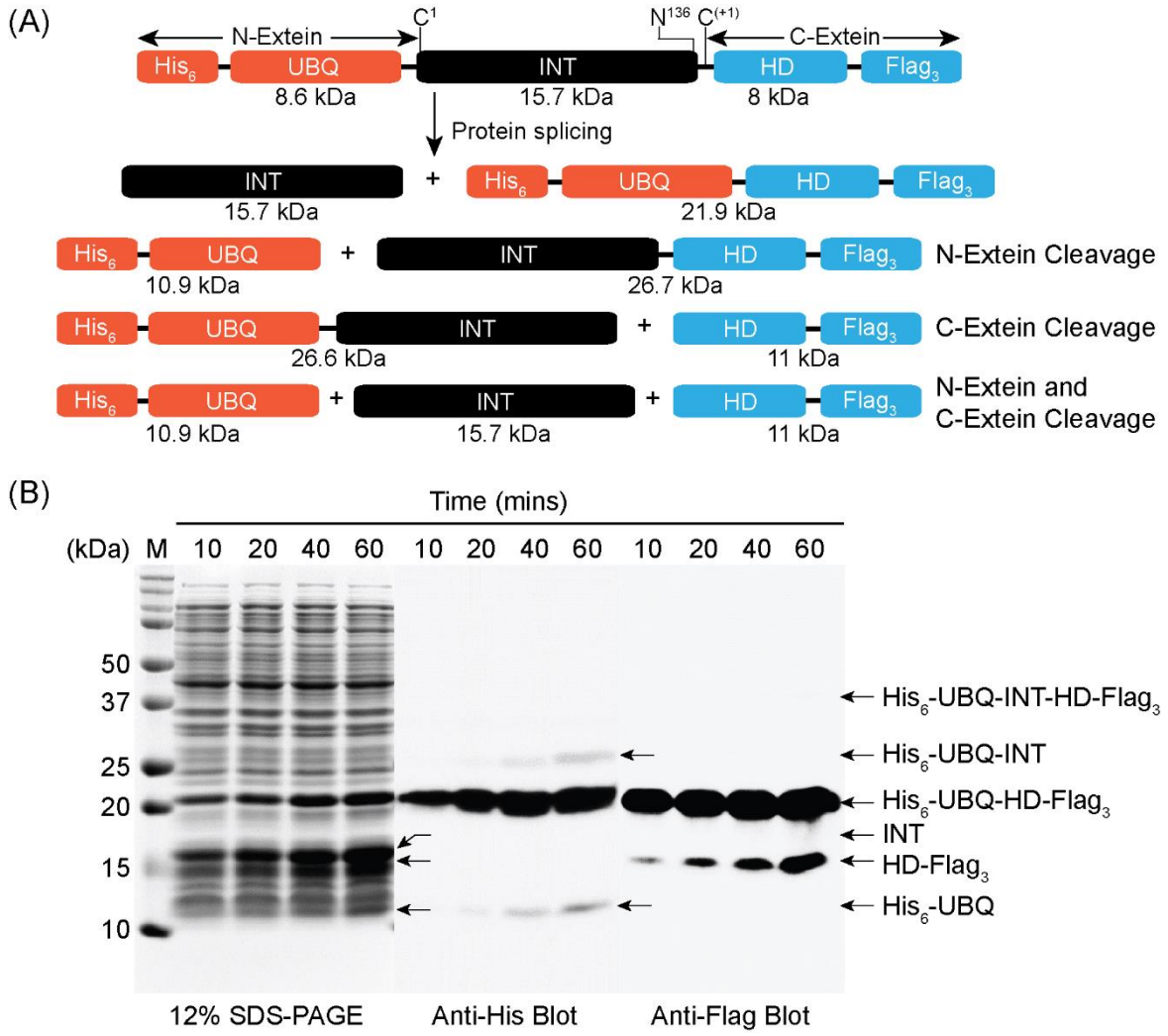


Figure 2 *Spl* DnaX mini intein is catalytically active

(A) Schematic representation of intein construct is shown where *Spl* DnaX mini-intein (INT) is fused with ubiquitin (UBQ) and homeodomain (HD) as N-extein and C-extein, respectively. The N- and C-terminus have His_6 and 3X-Flag tags, respectively. Positions of the three catalytically important residues Cys1, Asn136 and Cys(+1) are shown. The spliced products and cleavage products due to side reactions are also indicated with their respective molecular weights. **(B)** After induction, the *E. coli* cells were harvested at different times, as indicated, and analyzed by 12% SDS-PAGE. Each sample was run in triplicate and analyzed by coomassie blue staining, anti-His_6 and anti-Flag western blots. The splicing and cleavage reaction products are indicated by arrows.

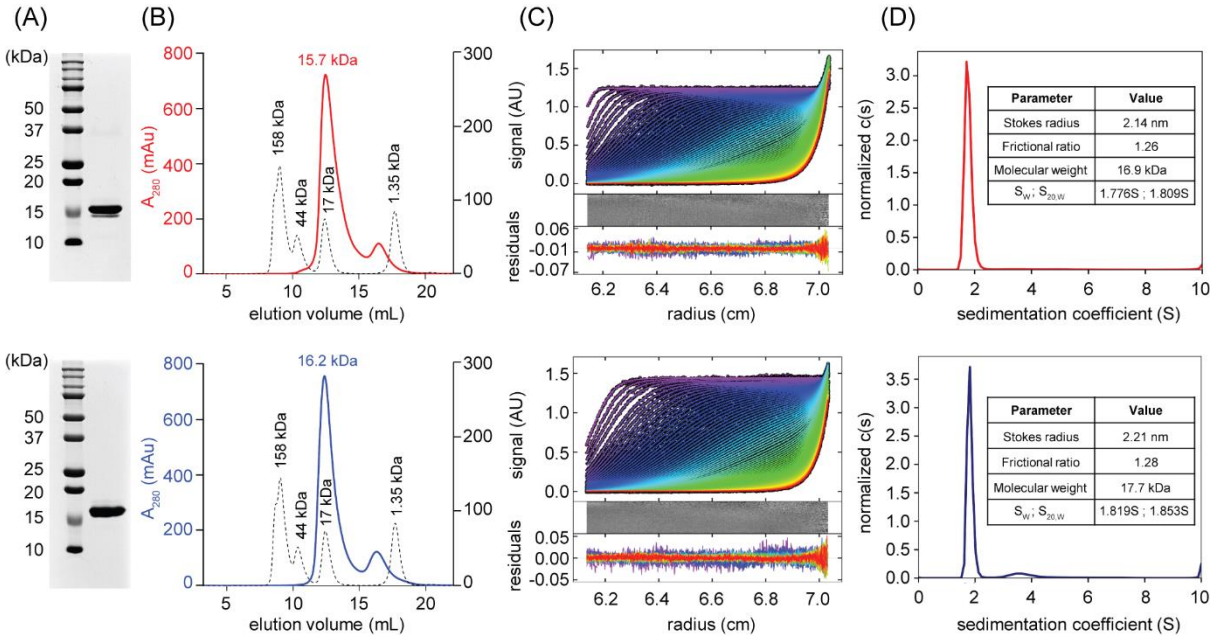


Figure 3 Expression and purification of *SpI* DnaX mini intein

(A) 12% SDS-PAGE of Intein^{WT} (top) and Intein$^{C1A/C(+1)A}$ (bottom) show high protein purity. **(B)** Analytical SEC profile shows that both Intein^{WT} (top) and Intein$^{C1A/C(+1)A}$ (bottom) eluted almost at the same volume. The dotted line is for protein standards. **(C)** SV-AUC statistical analysis of Intein^{WT} (top) and Intein$^{C1A/C(+1)A}$ (bottom) is shown. The upper panel in the figures shows the overlay of the experimental values (circle), and the fitted curve (solid line) with RMSD value of 0.007 for both constructs. The lower panel shows the corresponding residuals of the fitted data. **(D)** AUC distance distribution *c*(*S*) versus sedimentation coefficient (*S*) plot for Intein^{WT} (top) and Intein$^{C1A/C(+1)A}$ (bottom) generated a single sharp peak indicating a monodispersed and homogeneous protein sample.

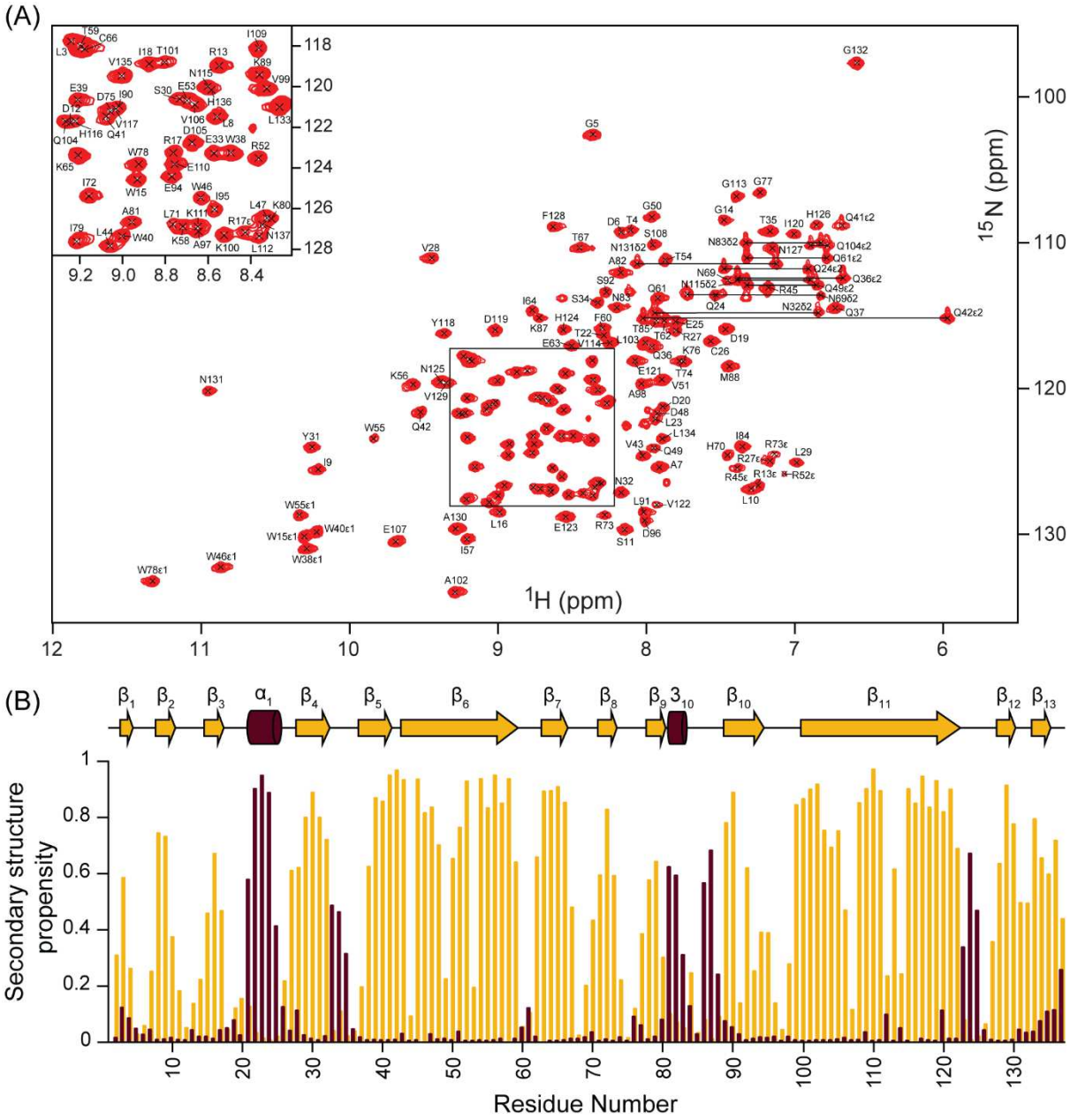


Figure 4 Backbone assignment and secondary structure prediction of *Spl* DnaX mini intein

(A) ^{15}N-^{1}H HSQC spectrum of Intein$^{C1A/C(+1)A}$ with the backbone assignments is shown. The *inset* shows the labelled peaks in the central crowded region of the spectrum. Sidechain resonances of Asn and Gln are connected with horizontal lines. **(B)** The secondary structure propensities of Intein$^{C1A/C(+1)A}$ is plotted for each residue. The propensity of the strand and helices are in orange and maroon, respectively. The resulting secondary structure is shown on the top with strands as arrows and helices as cylinders. The numbering starts with the N-extein residue Glu(-1) as 1.

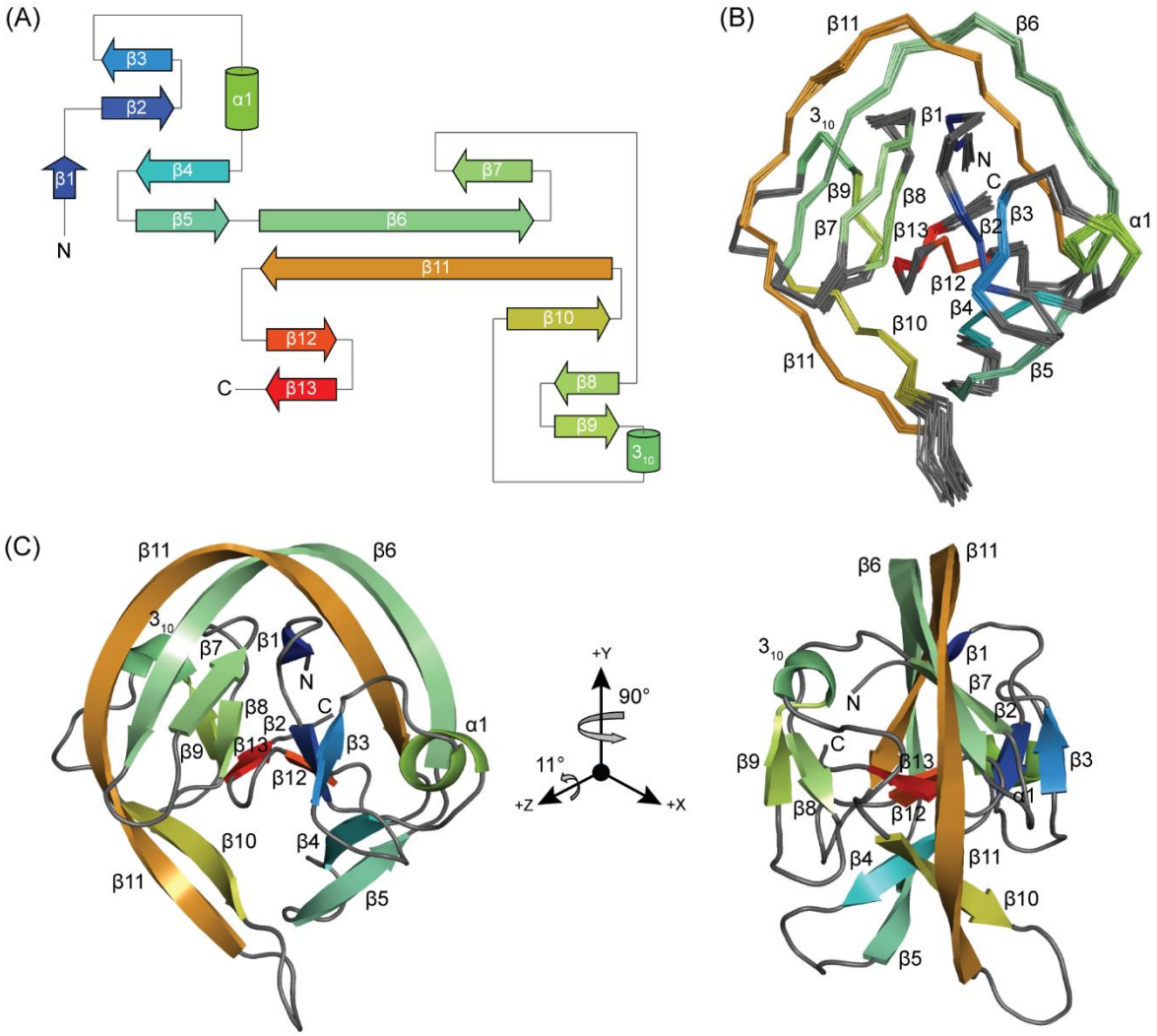


Figure 5 Structural ensembles of *Spl* DnaX mini intein

(A) The secondary structure topology map of *Spl* DnaX mini-intein is shown. **(B)** An ensemble showing C-alpha trace of 15 energy-minimized conformers representing the three-dimensional structure of *Spl* DnaX mini-intein, with a backbone atom RMSD of 0.27 Å. Regions of secondary structure are labeled. The N- and C-termini are indicated. **(C)** Ribbon representation of the lowest energy conformer of *Spl* DnaX mini-intein is shown.

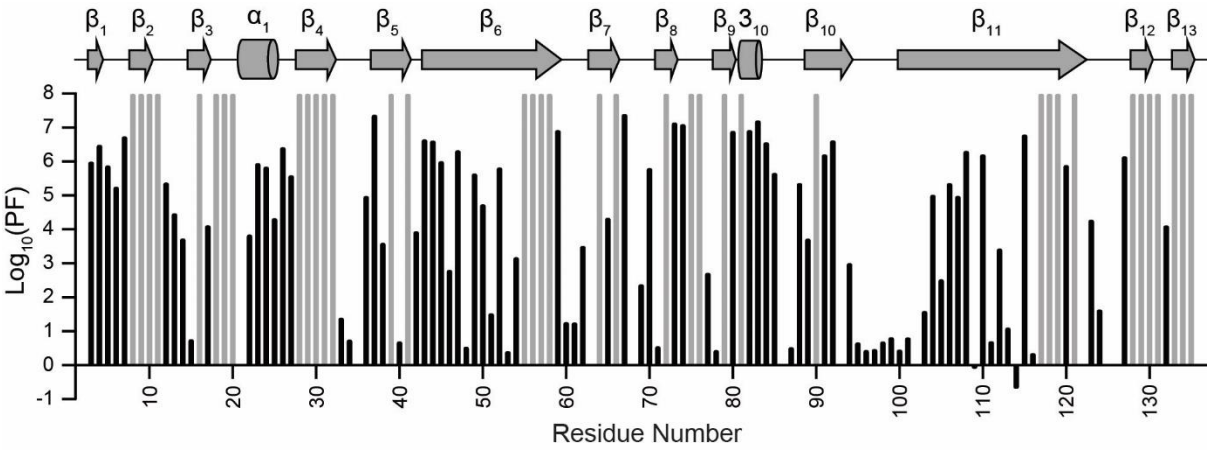


Figure 6 Amide HX studies show that *Spl* DnaX mini intein has a very stable core

Backbone amide HX studies demonstrate a wide range of stability. The slowest exchanging amides, which show no appreciable decay over a period of 110 days, are assigned a protection factor (PF) of 10^8 and colored in grey. The residues whose PFs were measured are shown in black. The missing bars correspond to prolines or residues that do not fit well. The secondary structure is shown on the top with arrows and cylinders for strands and helices, respectively. The numbering starts with the N-extein residue Glu(-1) as 1.

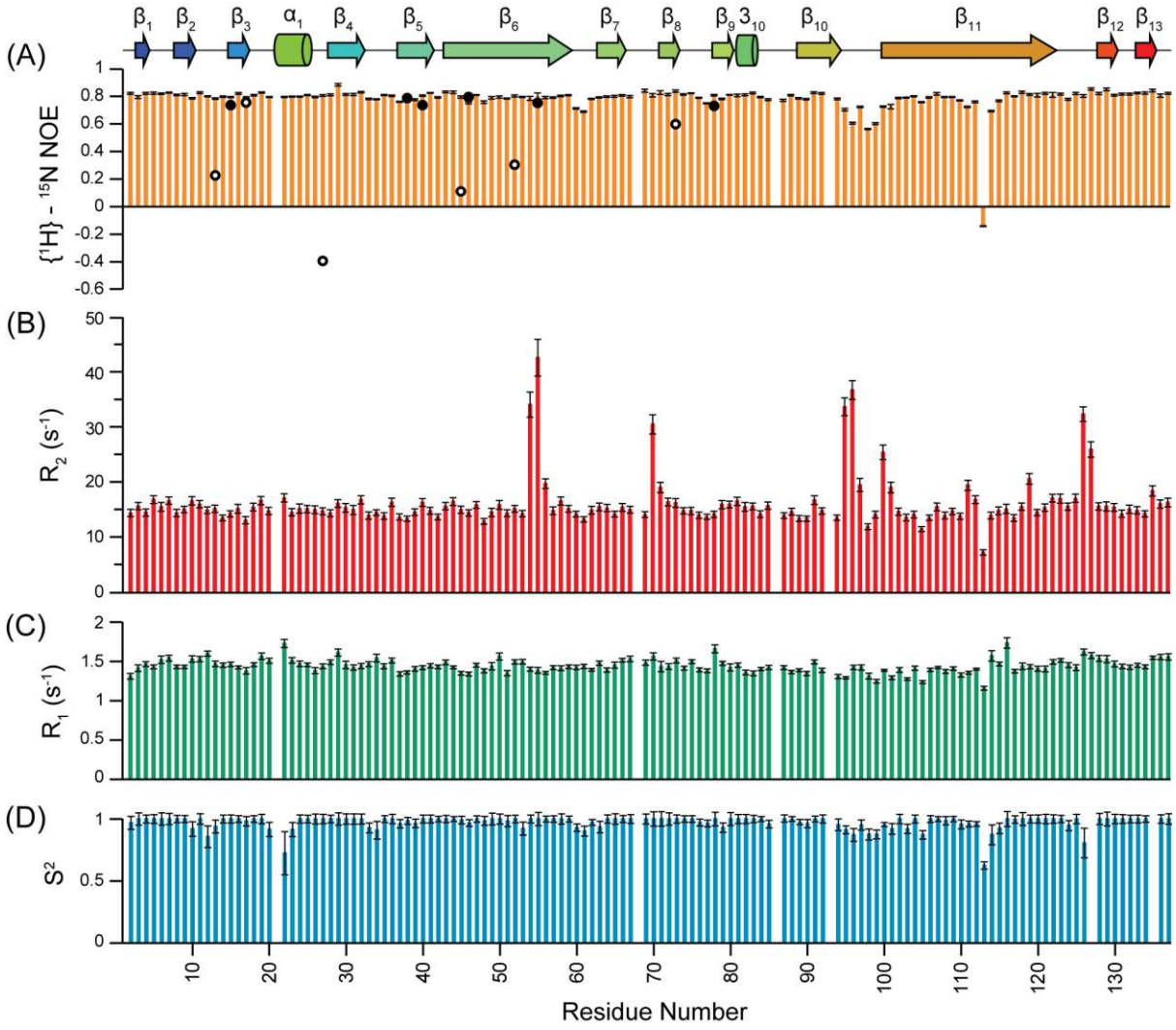


Figure 7 Fast nanosecond to picosecond timescale mobility of amides in *SpI* DnaX mini intein

(A) Heteronuclear {^{1}H}-^{15}N NOE, **(B)** transverse relaxation rate constants (R_2), **(C)** longitudinal relaxation rate constants (R_1) and **(D)** order parameter (S^2) values are plotted for each residue. The {^{1}H}-^{15}N NOE values of tryptophan (filled circle) and arginine (open circle) sidechains are also indicated. The secondary structure is shown on the top with arrows and cylinders for strands and helices, respectively. The numbering starts with the N-extein residue Glu(-1) as 1.

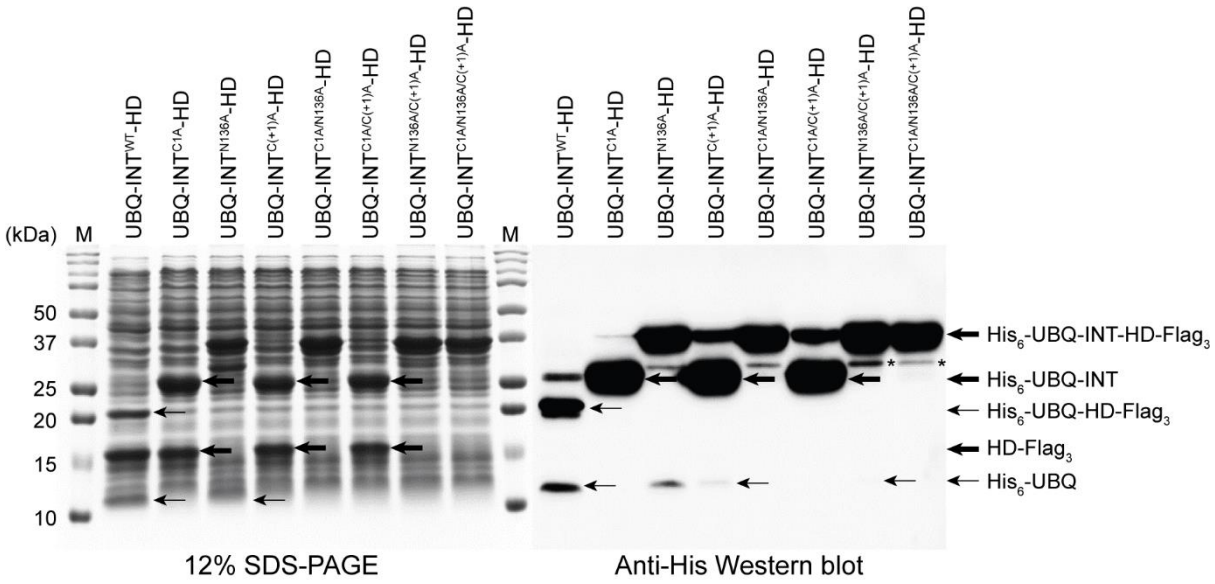
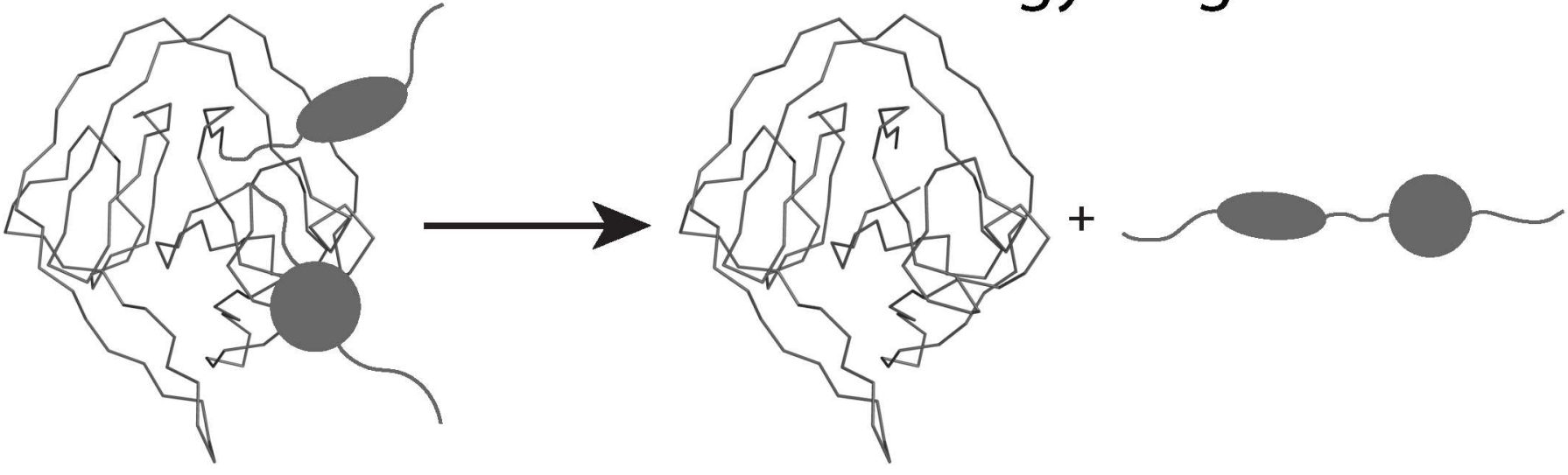


Figure 8 Comparative assay of *Spl* DnaX intein mutants

Left: 12% SDS-PAGE analysis of the wild type and mutants of *Spl* DnaX mini-intein. The cells of all the constructs were harvested at 1-hour post-induction. Right: Western blot using anti-His_6 antibody showing the spliced and side reaction cleavage products, as indicated by arrows. The asterisks denote the non-specific degradation products.

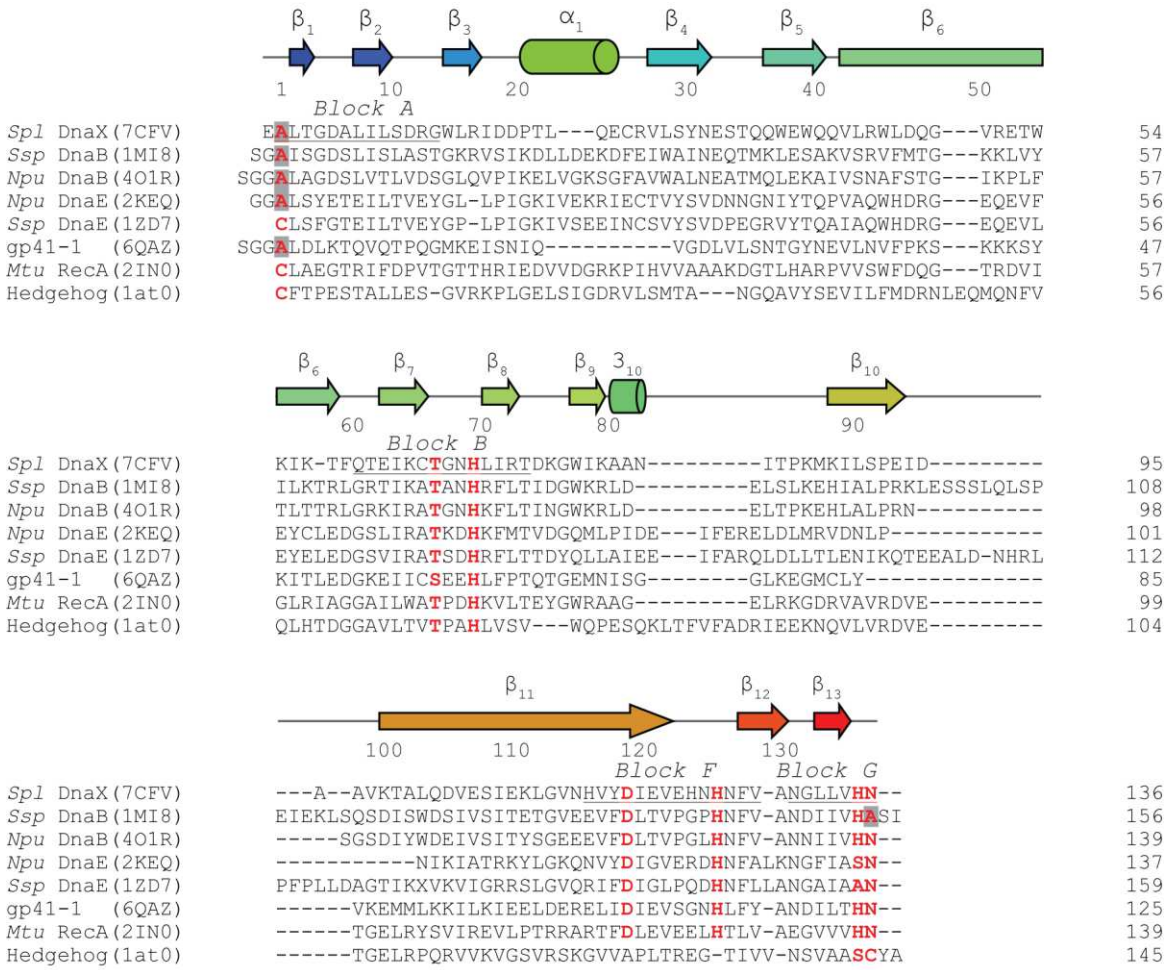


Supplementary table S1 Comparison of *Spl* DnaX mini-intein with other intein structures using the DALI server

Intein	Sequence Identity (%)	PDB Codes	Residues	Z-score[a]	RMSD[b] (Å)
Ssp DnaB	32	1MI8	141	18.9	1.9
	30	6FRE	154	18.2	2.1
	31	6FRH	160	17.9	2.2
Npu DnaB	31	4O1R	142	18.8	2.0
Npu DnaE	27	4KL5	140	16.2	2.1
Ssp DnaE	24	1ZD7	159	16.2	2.0
	25	1ZDE	160	16.0	2.2
gp41-1	29	6RIZ	127	16.6	2.0
	30	6QAZ	128	16.0	2.2
Mtu RecA	22	3IFJ	138	19.5	1.7
	22	2IMZ	143	19.3	1.9
	21	3IGD	131	19.2	1.8
	21	2IN9	139	19.1	1.8
	21	2IN8	139	19.1	1.8
	21	2IN0	139	19.1	1.8
	21	2L8L	139	17.5	2.0
Drosophila hedgehog	16	1AT0	145	17.2	2.2
	16	6TYY	146	16.9	2.2

[a]Z-scores[1] are computed by mean and standard deviation of pairwise structural comparison scores for proteins < 400 residues; Z-scores above 8 indicates very good structural superimpositions with probable homology between the two structures, whereas Z-scores above 20 means the two structures are definitely homologous.

[b]RMSD is calculated as the average deviation in distance between aligned alpha carbons in structural superimposition.



Supplementary figure S1 Structure-based sequence alignment of the *Spl* DnaX mini intein with other inteins

The sequence alignments were generated using Clustal Omega program[2] PDB accession codes are listed in brackets after each intein. In the sequence alignment, the conserved residues are shown in *red* and the mutated residues are highlighted in *gray*. The secondary structure of *Spl* DnaX mini-intein is shown at the top. The strands and helices are shown as arrows and cylinders, respectively. Residues in the four blocks A, B, F and G, are *underlined*. *Hyphens* (-) indicate gaps in the alignment.

REFERENCES

1 L. Holm and C. Sander, Protein Structure Comparison by Alignment of Distance Matrices, *J. Mol. Biol.*, 1993, **233**, 123–138.

2 J. D. Thompson, D. G. Higgins and T. J. Gibson, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 1994, **22**, 4673–4680.