Bioscience Reports

www.bioscirep.org

Bioscience Reports

Mutual synergistic protein folding in split intein

Yuchuan ZHENG*, Qin WU*, Chunyu WANG†, Min-qun XU‡ and Yangzhong LIU*[1]

*CAS Key Laboratory of Soft Matter Chemistry, Department of Chemistry, University of Science and Technology of China, 230026 Hefei, Anhui, People's Republic of China, †Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, U.S.A., and ‡New England Biolabs, 32 Tozer Road, Beverly, MA 01915, U.S.A.

Synopsis

Inteins are intervening protein sequences that undergo self-excision from a precursor protein with the concomitant ligation of the flanking polypeptides. Split inteins are expressed in two separated halves, and the recognition and association of two halves are the first crucial step for initiating *trans*-splicing. In the present study, we carried out the structural and thermodynamic analysis on the interaction of two halves of DnaE split intein from *Synechocystis* sp. PCC6803. Both isolated halves (I_N and I_C) are disordered and undergo conformational transition from disorder to order upon association. ITC (isothermal titration calorimetry) reveals that the highly favourable enthalpy change drives the association of the two halves, overcoming the unfavourable entropy change. The high flexibility of two fragments and the marked thermodynamic preference provide a robust association for the formation of the well-folded I_N/I_C complex, which is the basis for reconstituting the *trans*-splicing activity of DnaE split intein.

Key words: disordered protein, intein, NMR spectroscopy, protein folding.

Cite this article as: Zheng, Y., Wu, Q., Wang, C., Xu, M.-q. and Liu, Y. (2012) Mutual synergistic protein folding in split intein. Biosci. Rep. **32**, 433–442

INTRODUCTION

Protein splicing is a post-translational process in which intervening proteins, termed inteins, are self-excised from the precursor proteins, while the two flanking polypeptides (exteins) are seamlessly ligated together with a normal peptide bond [1]. Two types of protein splicing occur naturally: *cis*-splicing with inteins having continuous sequences or *trans*-splicing with inteins having split sequences. In comparison with *cis*-splicing, *trans*-splicing has one additional binding step, in which two fragments of split inteins recognize and associate with each other to reconstitute splicing activity, followed by the four steps in protein splicing mechanism (Figure 1) [2]. Protein *trans*-splicing can be easily controlled because each individual fragment can be prepared separately in an inactive form, and protein splicing occurs only after the reconstitution of the two complimentary fragments. Thus, split inteins have been used for various applications in protein engineering, such as segmental isotopic labelling, protein semi-synthesis and detection of protein–protein interactions [3–5].

The DnaE intein from the split *dnaE* gene of *Synechocystis* sp. PCC6803 is a naturally occurring split intein, which has *trans*-splicing activity both *in vivo* and *in vitro* [6,7]. The crystal structure of an artificially fused continuous DnaE intein has been reported to adopt a horseshoe-like three-dimensional structure, termed the HINT (hedgehog/intein) fold, although the attempt to co-crystallize the two fragments of DnaE split intein failed [8]. A fast association rate (2.8×10^{7} $M^{-1}\cdot s^{-1}$) with high affinity (K_d~36 nM) of DnaE split intein was determined by FRET (fluorescence resonance energy transfer) [9]. Furthermore, the electrostatic interactions between split intein halves have been postulated to play a critical role during the binding process of the split DnaE intein [9–11]. Although the protein splicing mechanism has been extensively studied [11–14], the recognition of two fragments of split inteins, which determines the reconstitution of *trans*-splicing activity, is still poorly understood at the atomic level. All reported structures of inteins, including artificially fused split inteins, share a common HINT fold. However, the artificially split protein fragments from continuous inteins are prone to misfold or aggregate in individual expression in *Escherichia coli* [15]. The solution behaviours of individual

Abbreviations used: CDF, cumulative distribution function; CH plot, charge hydropathy plot; DLS, dynamic light scattering; DTT, dithiothreitol; GB1, B1 domain of protein G; GdmCl, guanidinium chloride; HINT, hedgehog/intein; HSQC, heteronuclear single quantum correlation; IDP, intrinsically disordered protein; ITC, isothermal titration calorimetry; IPTG, isopropyl β-D-thiogalactoside; KSI, ketosteroid isomerase; LB, Luria–Bertani; Ni-NTA, Ni^{2+}-nitrilotriacetate; RBS, ribosome-binding site; SET, solubility enhancement tag; TEV, tobacco etch virus.

[1] To whom correspondence should be addressed (email liuyz@ustc.edu.cn).

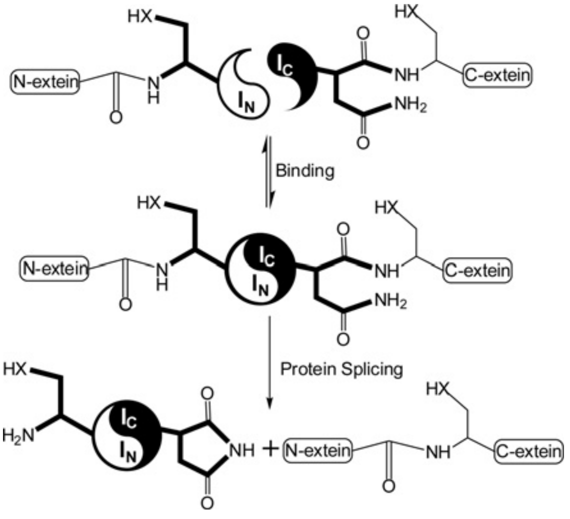

Figure 1 Schematic representation of protein *trans*-splicing
I_N and I_C represent the N- and C-terminal fragments of a split intein respectively. The two halves bind to each other non-covalently to initiate the splicing activity. The mature host proteins are produced by protein splicing while the split intein binary complexes are excised. X denotes an O or S atom.

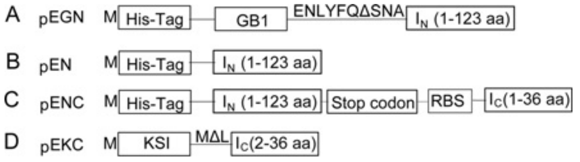
A pEGN M His-Tag GB1 ENLYFQΔSNA I_N (1-123 aa)
B pEN M His-Tag I_N (1-123 aa)
C pENC M His-Tag I_N (1-123 aa) Stop codon RBS I_C(1-36 aa)
D pEKC M KSI MΔL I_C(2-36 aa)

Figure 2 Schematic representation of the constructs used in the present study
(**A**) pEGN was used for the preparation of I_N fusion. The symbol Δ indicates the cleavage site between Q and S. (**B**) pEN was used for the expression of sole I_N. (**C**) pENC is the co-expression system for the expression of I_N and I_C. (**D**) pEKC was used for the expression of I_C fusion. The symbol Δ indicates the CNBr cleavage site. aa, amino acid.

halves from the naturally occurred split intein and their recognition have not been carefully studied.

To gain insight into the recognition of split inteins, we cloned and purified two halves of DnaE split intein, I_N containing three extein residues and I_C without extein. Our results show that, in the absence of the complementary partner, both I_N and I_C are unstructured under physiological conditions. The conformational changes occur upon association in an enthalpy-driven process. The inherent flexibilities with abundance of polar contacts provide an efficient approach to reconstitute protein *trans*-splicing.

EXPERIMENTAL

Plasmids

The gene encoding I_N (amino acids 1–123) was amplified by PCR from pMEB10 with the following primers: forward, 5′-TACTTCCAATCCAATGCGTGCCTGTCTTT-3′; and reverse, 5′-TTATCCACTTCCAATGCTATTTAATTGTCCCAG-3′. The PCR product was cloned into a modified pMCSG7 vector with an N-terminal His_6-GB1 (the B1 domain of protein G) tag as a SET (solubility enhancement tag) using the LIC (ligation-independent cloning) method. This construct is named pEGN for preparation of the N-terminal fragment of DnaE split intein (Figure 2A). The gene for His_6-I_N was amplified from pMEB10 via PCR using forward (5′- TGCGCTcatatgCACCATCATCATCACACTGCCTGTCTTT-3′, containing an NdeI site denoted in lower-case letters and a His_6 tag underlined) and reverse (5′-CGgaattcCTATTTAATTGTCCCAGCGT-3′, containing an EcoRI site denoted in lower-case letters and a stop codon underlined) oligonucleotide primers. The PCR product was cloned into the NdeI–EcoRI-digested pET-21b(+) vector (Novagen). This construct is called pEN for the expression of sole I_N without the SET (Figure 2B). The co-expression plasmid containing genes encoding His_6-I_N and I_C was constructed by two-step PCR using pEN described above according to the previously published method [16]. I_C and a second RBS (ribosome-binding site) were amplified with the I_C PCR products as the template described above using forward (5′-TAACTTTAAGAAGGAGATATACCATGGTTAAAGTTATCGGTCGT-3′, with RBS underlined) and reverse (5′- CCGctcgagCTAGTTAGCAGCGATAGC-3′, containing an XhoI site denoted in lower-case letters and a stop codon underlined) oligonucleotide primers. The second PCR was done with the first PCR product as the template using the forward (5′-CGgaattcAAATAATTTTGTTTAACTTTAAGAAGGAGATAT-3′, containing an EcoRI site denoted in lower case letters) oligonucleotide primer and the reverse primer remained unchanged. The resulting PCR products were digested and cloned into the EcoRI–XhoI-digested pEN. This construct is named pENC for the co-expression of I_N/I_C complex (Figure 2C). The gene for I_C (2–36 amino acid) was amplified from pKEC3 via PCR using forward (5′-CGAcagatgctgGTTAAAGTTATCGGTCG-3′, containing a AlwNI site denoted in lower case letters) and reverse (5′-CCGctcgagCTAGTTAGCAGCGATAGC-3′, containing a XhoI site denoted in lower case letters) oligonucleotide primers. The PCR product was cloned into the AlwNI-XhoI-digested pET-31b(+) vector (Novagen), a vector for high-level expression of peptide with an N-terminal KSI (ketosteroid isomerase) tag. This construct is named pEKC for preparation of the C-terminal fragment of DnaE split intein (Figure 2D). All clones were verified by DNA sequencing.

Protein expression and purification

N-terminal fragment (I_N)

Plasmid pEGN was introduced into *E. coli* ER2566 cells. The cells were grown in 1 litre of LB (Luria–Bertani) medium with 100 μg/ml ampicillin at 37 °C until the cells reached a D_{600} of 0.6–0.8, then the protein expression was induced by 0.5 mM

IPTG (isopropyl β-D-thiogalactoside) at 16 °C for 16 h. The cells were harvested by centrifugation at 4000 $\boldsymbol{g}$ for 20 min and resuspended in buffer A (50 mM Tris/HCl, pH 8.0, 200 mM NaCl, 5 mM 2-mercaptoethanol and 10 mM imidazole), and sonicated in an ice bath. Cell lysates were centrifuged at 22 324 $\boldsymbol{g}$ for 30 min at 4 °C, and then the supernatant was loaded on to a column of Ni-NTA (Ni^{2+}-nitrilotriacetate) resin (Qiagen) pre-equilibrated with buffer A. The fusion protein was eluted by buffer B (buffer A with 250 mM imidazole). The protein was concentrated and the imidazole was removed by ultrafiltration using a membrane with a 10-kDa molecular mass cut-off (Amicon). The His_6-GB1 tag was digested using TEV (tobacco etch virus) protease at 16 °C overnight in 50 mM Tris/HCl (pH 8.0), 100 mM NaCl, 0.5 mM EDTA and 1 mM DTT (dithiothreitol). After removal of EDTA and DTT by ultrafiltration with a 3-kDa molecular mass cut-off, the enzyme-digested products were purified again using Ni-NTA resin to remove His_6-GB1 tag, TEV protease and the undigested proteins. The I_N (containing three additional N-terminal residues (Ser-Asn-Ala) from TEV digestion) in the flow-through fraction was purified further by a Superdex-75 gel filtration column 16/60 (GE Healthcare) with eluent containing 20 mM Tris/HCl, pH 7.4, 100 mM NaCl, 1 mM EDTA and 5 mM 2-mercaptoethanol. The molecular mass of I_N was confirmed by SDS/PAGE (15 % gel) and MS. The concentration of I_N was determined by UV absorption at 280 nm with the theoretical molar absorption coefficient ($\varepsilon = 12950\ M^{-1} \cdot cm^{-1}$).

C-terminal fragment (I_C)

Plasmid pEKC was transformed into *E. coli* BL21 (DE3). The cells were grown in 1 litre of LB medium with 100 μg/ml ampicillin at 37 °C until the cells reached a D_{600} of 0.6–0.8, then the protein expression was induced by 0.8 mM IPTG at 37 °C for 5 h. I_C was prepared using method given in the literature [17] with the following modifications. After the CNBr digestion, the pH of the solution diluted 1:4 (v/v) with MiniQ water was adjusted to approximately 4.5 with ammonia to yield white precipitates containing KSI and I_C. I_C was released from the precipitate by MiniQ water treatment and further purified by RP-HPLC (reverse-phase HPLC) using a C_{18} kromasil column (10 mm×250 mm, 5 μm) at a flow rate of 3 ml/min with two-step linear gradient: 0–5 min: 10–30 % eluent B; 5–20 min: 30–100 % eluent B, where eluent A is 0.1 % TFA in Mini-Q water and eluent B is methanol. The purified I_C was freeze-dried. The molecular mass of I_C (containing one additional N-terminal leucine residue from CNBr cleavage) was confirmed by MS. The concentration of I_C was determined by the BCA (bicinchoninic acid) assay.

Co-expression system

The co-expression plasmid pENC was transformed into ER2566 cells. The cells were grown in 1 litre of LB medium with 100 μg/ml ampicillin at 37 °C. Protein expression was induced when cells reached a D_{600} of 0.6–0.8 with 0.1 mM IPTG at 37 °C for 3 h. Cells harvested from cultures were resuspended in buffer A and sonicated in an ice bath. The I_N/I_C complex was isolated from the supernatant on an Ni-NTA column and further purified by a Superdex-75 gel filtration column 16/60 (GE Healthcare) as described for the purification of I_N. The concentration of the I_N/I_C complex was also determined by UV absorption at 280 nm with the theoretical molar absorption coefficient ($\varepsilon = 12950\ M^{-1} \cdot cm^{-1}$).

N-terminal fragment without SETs

For expression of sole I_N, plasmid pEN was transformed into *E. coli* ER2566. The cells were grown in 4 ml LB medium with 100 μg/ml ampicillin at 37 °C. Protein expression was induced when cells reached a D_{600} of 0.6–0.8 with 0.1 mM IPTG at 37 °C for 3 h. The cells were harvested using 1.5 ml Eppendorf tube and resuspended using buffer A, and then sonicated in an ice bath. Cell lysates were centrifuged at 22 324 $\boldsymbol{g}$ for 10 min at 4 °C and the inclusion bodies were dissolved in 8 M urea. Both supernatant and inclusion bodies were analysed by gel electrophoresis. In order to compare the influence of I_C on the expression of I_N, both uninduced and induced cells harbouring the co-expression plasmid pENC were treated in the same manner as cells harbouring pEN.

^{15}N-Labelled proteins

The uniformly ^{15}N-labelled samples were prepared in the same procedure, except for replacing the LB medium with the M9 medium supplemented with $^{15}NH_4Cl$ as the sole nitrogen source. The samples of I_N and the I_N/I_C complex were concentrated by ultrafiltration and the buffer was exchanged to NMR buffer (100 mM NaCl, 3 mM DTT and 20 mM sodium phosphate buffer at pH 7.0). The freeze-dried I_C was dissolved directly in NMR buffer.

CD spectroscopy

Far-UV CD spectra were obtained with a Jasco J-810 apparatus equipped with a temperature-controlled water bath using quartz cuvette with a 1.0 mm path length at 25 °C from 190 to 260 nm in 50 mM sodium phosphate buffer (pH 7.4). Spectra were recorded with 15 μM I_N with buffer solution as the blank. All measurements were repeated three times. The protein denaturation experiments were performed by adding 0–6 M GdmCl (guanidinium chloride; 1 M increments) to protein samples. CD spectra of the thermal denaturation experiments were recorded at 25, 55, 75 and 95 °C respectively. The sample was allowed to stand for 5 min at each temperature prior to the measurements.

Gel-filtration chromatography

Gel filtration was performed on an AKTA purifier liquid chromatography system using a Superdex-75 10/300 GL prepacked column (GE Healthcare), which was pre-equilibrated with 20 mM Tris/HCl containing 100 mM NaCl (pH 8.0). The column was calibrated with RNase A (13.7 kDa, 16.4 Å; where 1 Å = 0.1 nM), chymotrypsinogen A (25.7 kDa, 20.9 Å), ovalbumin (44 kDa, 30.5 Å) and BSA (66.2 kDa, 35.5 Å). The results were analysed according to the method described previously [18].

DLS (dynamic light scattering)

DLS measurements were performed at 25 °C on a DynaPro 99 instrument equipped with a temperature-controlled microsampler (Protein Solutions) at a laser wavelength of 824.3 nm. The purified I_N was centrifuged at 22 324 g for 10 min at 4 °C prior to measurements. Data were recorded using 0.5 mg/ml I_N in 20 mM Tris/HCl (pH 7.4) containing 100 mM NaCl, 1 mM DTT and analysed using the DYNAMICS V6.0 software from Protein Solutions.

Bioinformatic analysis

The amino acid composition of I_N was analysed using the composition profiler software [19], which includes the Disprot 3.4 and PDB_Select_25 datasets. The disorder prediction of I_N was performed using PONDR® software with VL-XT algorithm (http://www.pondr.com/), and then CDF (cumulative distribution function) analysis of the output of PONDR® was performed to provide the distribution of prediction scores and a linear boundary for distinguishing ordered and disordered proteins. The CH plot (charge hydropathy plot) was also done using PONDR®.

NMR spectroscopy

All ^1H-^{15}N HSQC (heteronuclear single quantum correlation) spectra were recorded at 298 K on 500 MHz Bruker spectrometer equipped with a cryoprobe. All uniformly ^{15}N-labelled protein samples were prepared in NMR buffer (100 mM NaCl, 3 mM DTT, 20 mM sodium phosphate buffer at pH 7.0) with 10 % ^2H_2O. Spectra were processed and analysed using NMR Pipe [20] and Sparky (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco, CA, U.S.A.).

Limited proteolysis

Limited proteolysis was carried out on I_N and the co-expresed I_N/I_C complex at 25 °C with trypsin in 20 mM Tris/HCl, 1 mM DTT, 0.5 mM EDTA and 100 mM NaCl (pH 8.0). The enzyme to substrate ratio was 1:1000 (w/w). Aliquots were taken from the proteolysis solution at the various reaction times (0, 2, 4, 8, 16, 32 and 64 min), and the reaction was stopped by adding SDS loading buffer and boiling for 5 min. Cleavage products were resolved by Tris-tricine gel electrophoresis (15 % gel) and visualized by staining with Coomassie Brilliant Blue R250.

ITC (isothermal titration calorimetry)

ITC experiments were performed on a Microcal ITC 200 (GE Healthcare) by titrating 150 μM I_C into 20 μM I_N at 25 °C in 20 mM sodium phosphate buffer (pH 7.4) and 100 mM NaCl. The heat of dilution was measured by titrating I_C into blank buffer, and the net binding heat was obtained by subtracting the dilution heat from the apparent reaction heat. Data were analysed using Origin ITC software.

RESULTS AND DISCUSSION

Expression of I_N and I_C

To study the interaction of two DnaE split intein fragments (I_N and I_C) and the properties of these fragments in solution, we first constructed a plasmid to produce I_N with an N-terminal His_6 tag, but this fusion protein was expressed mainly in insoluble inclusion bodies at 16 °C in *E. coli*. To improve high-level expression of soluble I_N, we next constructed a plasmid to produce I_N with an N-terminal His-tagged GB1 (6.2 kDa) as a SET and a TEV protease recognition site between GB1 and I_N. The fusion protein (His_6-GB1-I_N) was purified by Ni-NTA affinity chromatography from supernatant and the fusion tag was removed by digestion with TEV protease. After protease digestion, I_N was purified further by Ni-NTA affinity chromatography, followed by gel filtration. The purified I_N was readily degraded at room temperature (25 °C) by residual protease contaminants in the protein preparation. The degradation can be effectively inhibited with 2 mM protease inhibitor PMSF (results not shown). Therefore freshly prepared I_N was used in the present study.

I_C was prepared using the pET-31b(+) vector (Novagen) for the high-level expression of peptide. The protein was expressed with an insoluble expression tag (KSI tag) at the N-terminus in order to protect the peptide from proteolytic degradation *in vivo*. The KSI-I_C fusion was extracted from inclusion bodies using 8 M urea. The fusion protein was precipitated after removing urea by dialysis against ultra-pure water, and then freeze-dried. The KSI tag was removed by CNBr digestion in 88 % formic acid. I_C was purified using RP-HPLC. The identities of both I_N and I_C were confirmed by MS.

Isolated fragments of split intein are disordered

Based on the intein structures solved so far, all inteins possess the β-rich HINT fold, which brings the N- and C-terminal splicing junctions close in space. Such a three-dimensional arrangement must play a crucial role in protein splicing. CD spectroscopy is the most commonly used method to determine the extent of secondary structure with the spectroscopic signatures of the α-helices (negative bands at 208 and 222 nm) and β-sheets (a negative minimum at 218 nm and a positive maximum at 195 nm), whereas the random coils show a minimum near 198 nm. CD spectra showed that I_N possesses a nearly unstructured conformation, characterized by a single negative minimum at 203 nm (Figure 3A, black curve). As the artificially fused Ssp DnaE intein showed a well-ordered structure, this observation suggests that the short peptide sequence of I_C plays an important role in assisting the folding of I_N. Meanwhile, the shoulder approximately 220–230 nm in CD spectra of I_N indicates that this disordered protein contains some residual secondary structure [21]. The chemical denaturation experiments confirmed the presence of the residual secondary structure, which were gradually lost with the addition of GdmCl (Figure 3B). It has been proposed that the residual structure in unstructured proteins or protein domains serves as a primary contact site and guides correct protein

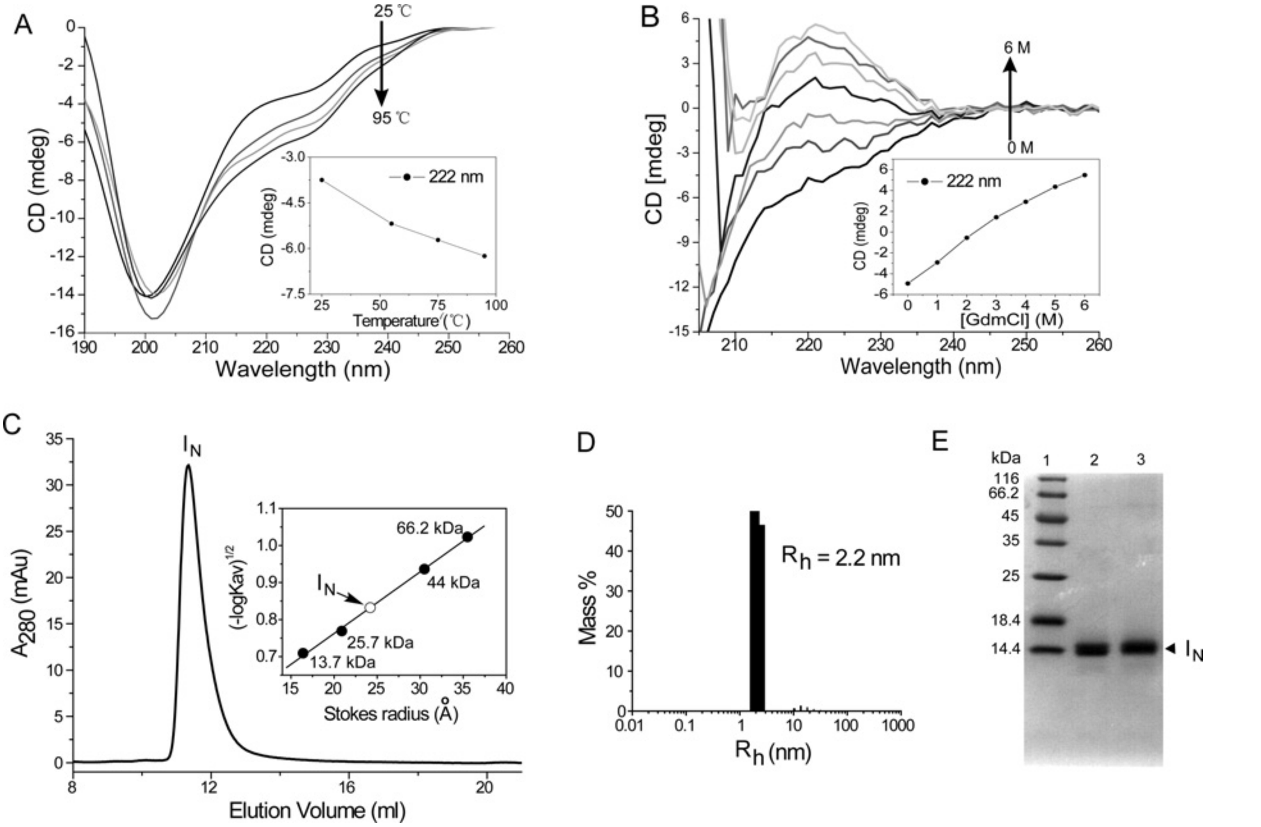

Figure 3 ***I_N* is intrinsically disordered *in vitro***

(**A**) Far-UV CD spectra of 15 μM I_N in 20 mM sodium phosphate buffer (pH 7.4). Spectra were recorded at 25, 55, 75 and 95 °C respectively. The ellipticity (mdeg) at 222 nm was plotted against temperature in the inset. (**B**) The chemical denaturation of I_N. Far-UV CD spectra were recorded with 15 μM I_N in various concentrations of GdmCl. The inset shows a continuous increase of ellipticity at 222 nm with the concentration of GdmCl. (**C**) Determination of Stokes radius by gel filtration. The panel shows the elution profile of I_N on a Superdex-75 10/300 GL prepacked column (GE Healthcare) at a flow rate of 0.6 ml/min. The inset shows a plot of Stokes radius against $(-\log K_{av})^{1/2}$, which used to determine Stokes radius of I_N (○). The linear calibration line was generated using four globular protein standards (●). (**D**) The DLS profile of I_N. The average hydrodynamic radius (R_h) was 2.2 nm as determined using 0.5 mg/ml I_N at 25 °C in 20 mM Tris/HCl (pH 7.4) containing 100 mM NaCl and 1 mM DTT. (**E**) Heating stability of I_N. Samples were analysed by SDS/PAGE (15 % gel), and visualized by Coomassie Blue staining. Lane 1, protein markers; lane 2, non-boiled sample; lane 3, boiled I_N, of which the protein sample was incubated in boiling water for 10 min followed by centrifugation at 22 324 ***g*** for 10 min. The supernatant was taken for the gel analysis, showing no obvious difference from the non-boiled sample.

folding by limiting the conformational space in protein–protein recognition [22]. Therefore the residual structure in I_N can be helpful in the association of two halves of DnaE split intein.

To test the stability of the residual secondary structure in I_N, the CD spectra were recorded from 25 to 95 °C. Interestingly, the CD spectra of I_N exhibit very little change over the temperature range, in which the ellipticity (mDeg) at 222 nm slightly decreases with temperature increase (Figure 3A). The temperature independent CD spectra are consistent with IDPs (intrinsically disordered proteins) [23]. Disordered proteins typically exhibit abnormally larger hydrodynamic radius (Stokes radius) than the globular proteins with the same molecular mass [24]. To test whether this is true for I_N, the hydrodynamic radius of I_N was first measured by gel filtration. The results showed that I_N (14.2 kDa) has a larger hydrodynamic radius (24 Å) than the corresponding molecular mass of globular protein {19 Å, calculated by the equation: $\log(R^N{}_S) = -0.204 + 0.357 \times \log$ M, where $R^N{}_S$ is the expected Stokes radius of globular proteins [23]} (Figure 3C). Meanwhile, the hydrodynamic radius of I_N determined by DLS gave a very similar value (22 Å) to the gel filtration result (Figure 3D). Furthermore, gel electrophoresis showed that the unusual solubility during boiling protects I_N from insoluble aggregates, which also supports the natively unfolded state of I_N (Figure 3E) [24].

For both I_N and I_C, two-dimensional ^{1}H-^{15}N HSQC NMR spectra were recorded on the uniformly ^{15}N-labelled protein samples. Results showed very poor dispersion for both I_N (Figure 4A, blue) and I_C (Figure 4B, blue) in the isolated form, which demonstrated the disordered state of these two fragments. Taken together, these results suggest that both I_N and I_C in separated form are in disordered state.

It has been reported that the IDPs typically have more disorder-promoting residues and less order-promoting residues, based on the bioinformatics analysis [25]. However, the amino acid

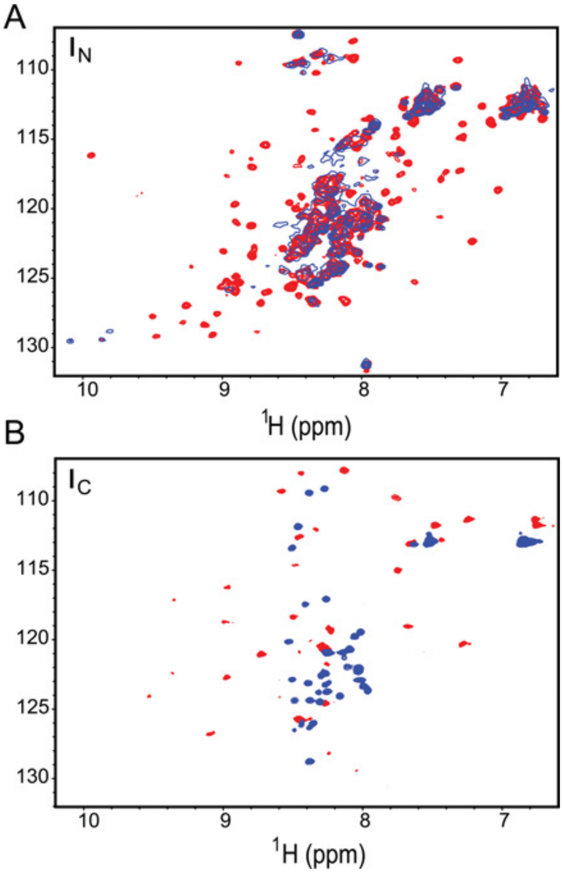

Figure 4 Both I_N and I_C undergo disorder-to-order transitions upon the association with each other

(**A**) Overlay of the ^{1}H-^{15}N HSQC spectra of uniformly ^{15}N-labelled I_N alone (blue) and in the presence of excess unlabelled I_C (red). (**B**) Overlay of the ^{1}H-^{15}N HSQC spectra of uniformly ^{15}N-labelled I_C alone (blue) and in the presence of excess unlabelled I_N (red).

composition analysis showed that I_N does not belong to the class of IDPs compared with the Disprot 3.4 dataset (Figure 5A). This result suggested that I_N has the potential to adopt an ordered structure. The intrinsic disorder propensity of I_N was further analysed by a neural network-based predictor PONDR® (http://www.pondr.com/) using the VL-XT algorithm with default parameters. The CDF analysis of the output of PONDR® has been used to provide the distribution of prediction scores and a linear boundary for distinguishing ordered and disordered proteins [26]. The CDF curve showed that I_N is located in the range of ordered proteins (Figure 5B). In addition, the CH plot, a plot of the normalized Kyte–Doolittle hydropathy against the mean net charge to classify ordered and disordered proteins [27], also located I_N in the ordered region (Figure 5C). These theoretical predictions consistently indicate that I_N has strong potential to form an ordered structure, although it demonstrated a disordered state experimentally. Therefore these results suggest that the disordered I_N is ready to fold into a well-structured protein in the association with its complementary partner.

Folding upon interaction

Although both I_N and I_C exhibit unfolded states in the isolated form, the X-ray crystal structure indicates that the artificially fused DnaE intein possesses a HINT structure. This fold is shared by all reported structures of inteins, so that it is very likely also present in split inteins after the association of two fragments, although no structures of split intein complex have been reported. Disordered proteins are usually involved in folding upon binding [28]. To verify the conformational change of I_N and I_C after association, ^{1}H-^{15}N HSQC spectra were recorded on ^{15}N-labelled I_N in the absence and presence of non-labelled I_C. The spectrum of I_N shows a significant improvement in signal dispersion after adding I_C (Figure 4A, red). A similar dispersion improvement was also observed on the ^{15}N-labelled I_C upon adding I_N (Figure 4B, red). This result clearly indicates the disorder-to-order transition of I_N upon binding to I_C, and vice versa. Therefore, it can be concluded that the disordered structure of the isolated fragments of I_N and I_C is due to the split of the continuous DnaE intein, and the mutual synergistic folding occurs during the association of I_N and I_C, which is a vital step for protein *trans*-splicing.

Limited proteolysis

It has been reported that the disordered proteins are more easily digested by proteases because the residues are largely exposed to solvent. On the contrary, ordered proteins are more resistant to proteases as the compact structures protect cleavage sites from protease access [29]. Therefore the limited proteolysis can be used to probe the conformational characteristics of proteins. Here, the protease sensitivity of I_N and I_N/I_C protein complex were examined using trypsin. Tris-tricine gel electrophoresis (15 % gel) was used to monitor the time course of protease digestion. The results show that I_N is prone to digestion as an individual fragment (Figure 6, upper panel); however, the protein gained great protease resistance in the co-expressed binary complex (Figure 6, lower panel). This result further confirmed that the disordered protein of I_N forms a folded three-dimensional structure in the binary complex.

I_C assists the folding of I_N *in vivo*

We next examined the association of two halves of DnaE intein *in vivo* using a co-expression system. The plasmid carrying two different genes, encoding I_N and I_C with two RBSs, was constructed according to the approach reported previously [16]. By transformation of the plasmid into *E. coli* ER2566 strain, both I_N and I_C were co-expressed simultaneously. In contrast with the expression of I_N alone, which mainly yielded the protein in insoluble inclusion bodies in *E. coli*, the co-expression system produced the soluble I_N/I_C complex in the supernatant (Figure 7A). The formation of insoluble I_N is likely attributed to the disordered nature of I_N, whereas the co-expression system generated the well-structured protein complex in the supernatant. It is well known that the newly synthesized proteins can escape protein misfolding and aggregation by folding modulators, such as molecular chaperones [30–32]. Thus, the co-expression data demonstrate that the expressed I_C can associate with expressed I_N within *E. coli* and assist the folding of I_N *in vivo*.

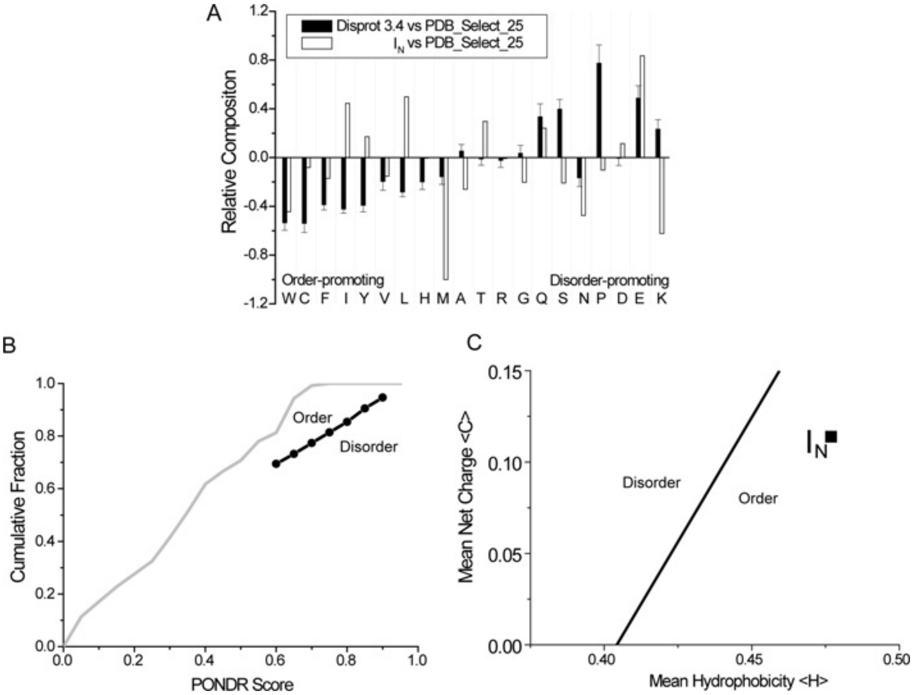

Figure 5 Bioinformatic analysis of I_N
(**A**) Comparisons of amino acid compositions of I_N (white bars) and Disprot 3.4 (black bars) with PDB_Select_25 datasets using Composition Profiler. The values of amino acids given according to Vihinen's flexibility scale represent enrichment (positive) or depletion (negative) relative to PDB_Select_25 dataset. I_N has less disorder-promoting amino acids but one residue (E, $P < 0.05$), and more order promoting amino acids than Disprot 3.4 dataset. (**B**) CDF analysis of I_N. The black line represents the boundary between order (above) and disorder (below). I_N located in the order region is shown as a grey curve. (**C**) The CH plot of I_N. A linear boundary (solid) ($\langle R\rangle = 2.785\langle H\rangle - 1.151$) separates the CH plot into disordered (above) and ordered (below) regions. I_N (■) is located at the ordered region.

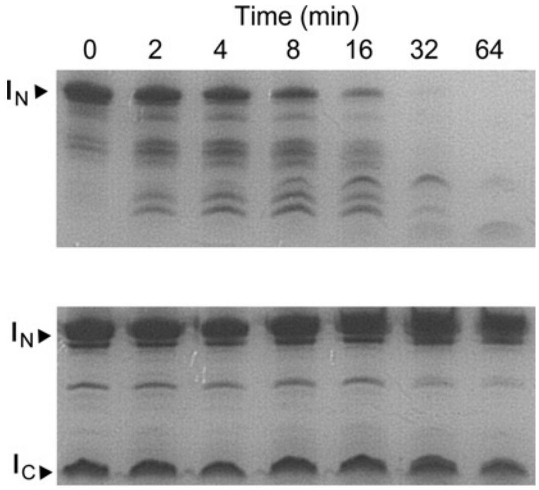

Figure 6 Limited proteolysis of purified I_N (upper panel) and I_N/I_C complex (lower panel) with trypsin
Reaction time was labelled in figure. The enzyme to substrate ratio was 1:1000 (w/w). Cleavage products were resolved by 15% Tris-tricine gel electrophoresis and visualized by staining with Coomassie Brilliant Blue R250.

To characterize further the I_N/I_C complex, the co-expression product was isolated using Ni-NTA affinity chromatography, and further purified by gel-filtration chromatography. The collected fractions were subjected to Tricine-SDS/PAGE (15% gel) and two protein bands corresponding to I_N and I_C were observed (Figure 7B). This result indicates that the I_N/I_C complex is highly stable during the chromatography processes. To verify the conformation of the binary complex, the ^{1}H-^{15}N HSQC NMR spectrum was recorded on the ^{15}N-labelled sample from the co-expression system. The well-dispersed backbone amide resonances provide solid evidence that the I_N/I_C complex is well folded (Figure 7C). Gel filtration clearly showed that the folded I_N/I_C complex has a smaller Stokes radius than disordered I_N (Figure 7D). This result reveals that the two fragments of split inteins can recognize and associate with each other *in vivo* to form a well-defined three-dimensional structure, which is essential for splicing activity. Similar results have also been observed in the split DnaE intein from *Nostoc punctiforme* (results not shown).

Thermodynamics of interaction

The thermodynamic parameters of the interaction between the two fragments of split intein were quantitatively measured by ITC. The titration of I_C into I_N was an exothermic process, illustrated by negative peaks in Figure 8. A non-linear best-fit binding isotherm yielded a dissociation constant (K_d) of 33 nM (Figure 8). This datum agrees with the previously reported result from the

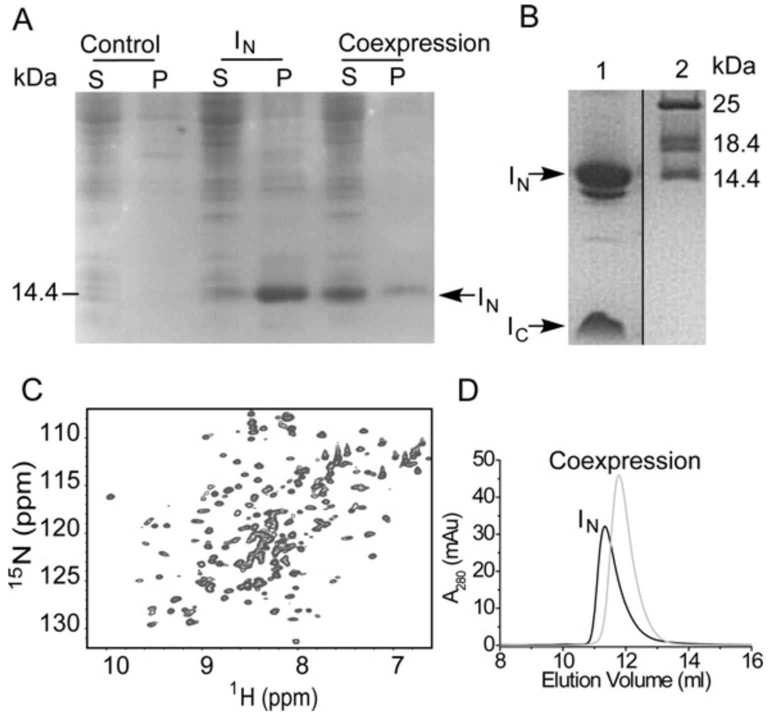

Figure 7 **Characterization of the I_N/I_C complex produced by co-expression**
(**A**) Recombinant expression of I_N alone and co-expression of I_N with I_C. The supernatant (S) and the pellet (P) from cell lysis were analysed by SDS/PAGE (15 % gel) and visualized by staining with Coomassie Brilliant Blue R250. The M_r of I_C is too small to be visualized on this gel. (**B**) The I_N/I_C complex purified from gel-filtration was resolved by 15 % Tris-tricine gel electrophoresis and visualized by staining with Coomassie Brilliant Blue R250. Lane 1 is co-expression products, lane 2 is protein markers. The bands corresponding to I_N and I_C are labelled in the Figure. (**C**) The ^{1}H-^{15}N HSQC NMR spectrum of the uniformly ^{15}N-labelled I_N/I_C complex from co-expression. (**D**) Gel filtration profiles of I_N alone and co-expression of I_N/I_C complex.

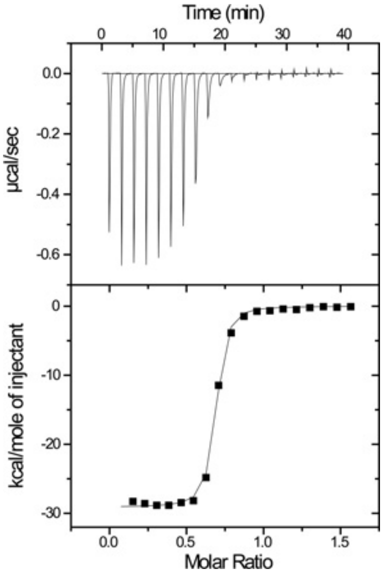

Figure 8 ITC profile of the titration of 150 μM I_C into 20 μM I_N
The top panel shows the raw data of the net binding isotherm, and the bottom panel shows the best fit curve (solid line) of the integrated enthalpy change (dot) to a one-site binding model using non-linear least-squares algorithm.

interaction of extein-containing I_N and I_C fragments [GFP (green fluorescent protein) as N-extein and organic fluorophore Texas Red as C-extein] measured by fluorescence [9]. The thermodynamic parameters of enthalpy change ($\Delta H = -29.1$ kcal/mol) and entropy change (TΔS $= -18.9$ kcal/mol) indicate that the binding is driven by enthalpy with a very unfavourable entropic cost. This phenomenon is known as the enthalpy–entropy compensation [33], which is often associated with folding upon binding. The favourable enthalpy change arises from a large number of polar contacts between I_N and I_C. The unfavourable entropy change is likely due to the loss of conformational freedom caused by the disorder-to-order transition during the association [34].

It is well known that the folded structure is a prerequisite for most proteins to perform various biological functions. However, accumulated evidence in the last decade showed that disordered proteins are involved in many cellular processes, such as transcriptional regulation and signal transduction [35]. On the other hand, disordered proteins could undergo conformational transition from disorder to well-folded structure upon binding to their partners and then carry out their specific biological functions [29,36–38], and the availability of disordered proteins are tightly controlled in cells [39]. Protein recognition and association can be through different mechanisms, including lock-and-key, induced-fit, conformational-selection and fly-casting.

The fly-casting mechanism suggests that the disordered proteins provide a greater capture radius for speeding up association followed by protein folding [40]. This mechanism may be applicable to the association of two split intein halves, which results in efficient binding and reconstitution of protein splicing activity.

Conclusions

In summary, the results of the present study reveal that both halves of DnaE split intein are unfolded, and undergo large conformational changes upon the association of two fragments. The mutual synergistic folding is featured by the highly favourable binding enthalpy change and highly unfavourable entropy change. The disordered nature of split intein halves, and the polar interactions between the two halves pave the way for reconstituting protein *trans*-splicing activity. This work provides insights into protein recognition of two halves of split inteins, which is the first crucial step for initiating protein *trans*-splicing.

AUTHOR CONTRIBUTION

Yuchuan Zheng and Yangzhong Liu designed the research and co-wrote the paper, Yuchuan Zheng and Qin Wu performed the research, and Chunyu Wang and Min-qun Xu provided technical support.

FUNDING

This work was supported by the National Science Foundation of China [grant number 20873135], the National Basic Research Program of China [973 Program, grant number 2009CB918804], the China Postdoctoral Science Foundation [grant number 2012M511408], and the National Institutes of Health [grant number GM081408 (to C.W.)].

REFERENCES

1 Noren, C. J., Wang, J. and Perler, F. B. (2000) Dissecting the chemistry of protein splicing and its applications. Angew. Chem. Int. Ed. Engl. **39**, 450–466

2 Du, Z., Shemella, P. T., Liu, Y., McCallum, S. A., Pereira, B., Nayak, S. K., Belfort, G., Belfort, M. and Wang, C. (2009) Highly conserved histidine plays a dual catalytic role in protein splicing: a pK_a shift mechanism. J. Am. Chem. Soc. **131**, 11581–11589

3 Zuger, S. and Iwai, H. (2005) Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. Nat. Biotechnol. **23**, 736–740

4 Mootz, H. D. (2009) Split inteins as versatile tools for protein semisynthesis. ChemBioChem **10**, 2579–2589

5 Kanno, A., Ozawa, T. and Umezawa, Y. (2006) Intein-mediated reporter gene assay for detecting protein-protein interactions in living mammalian cells. Anal. Chem. **78**, 556–560

6 Wu, H., Hu, Z. and Liu, X. Q. (1998) Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803. Proc. Natl Acad. Sci. U.S.A. **95**, 9226–9231

7 Martin, D. D., Xu, M. Q. and Evans, Jr, T.C. (2001) Characterization of a naturally occurring trans-splicing intein from *Synechocystis* sp. PCC6803. Biochemistry **40**, 1393–1402

8 Sun, P., Ye, S., Ferrandon, S., Evans, T. C., Xu, M. Q. and Rao, Z. (2005) 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. **353**, 1093–1105

9 Shi, J. and Muir, T. W. (2005) Development of a tandem protein trans-splicing system based on native and engineered split inteins. J. Am. Chem. Soc. **127**, 6198–6206

10 Dassa, B., Amitai, G., Caspi, J., Schueler-Furman, O. and Pietrokovski, S. (2007) Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations. Biochemistry **46**, 322–330

11 Shah, N. H., Vila-Perello, M. and Muir, T. W. (2011) Kinetic control of one-pot trans-splicing reactions by using a wild-type and designed split intein. Angew. Chem. Int. Ed. Engl. **50**, 6511–6515

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

13 Evans, T.J.T. and Xu, M. Q. (2002) Mechanistic and kinetic considerations of protein splicing. Chem. Rev. **102**, 4869–4884

14 Du, Z., Zheng, Y., Patterson, M., Liu, Y. and Wang, C. (2011) pK_a coupling at the intein active site: implications for the coordination mechanism of protein splicing with a conserved aspartate. J. Am. Chem. Soc. **133**, 10275–10282

15 Brenzel, S., Kurpiers, T. and Mootz, H. D. (2006) Engineering artificially split inteins for applications in protein chemistry: biochemical characterization of the split Ssp DnaB intein and comparison to the split Sce VMA intein. Biochemistry **45**, 1571–1578

16 Strong, M., Sawaya, M. R., Wang, S., Phillips, M., Cascio, D. and Eisenberg, D. (2006) Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. Proc. Natl Acad. Sci. U.S.A. **103**, 8060–8065

17 Zhao, D. X., Ding, Z. C., Liu, Y. Q. and Huang, Z. X. (2007) Overexpression and purification of single zinc finger peptides of human zinc finger protein ZNF191. Protein Expr. Purif. **53**, 232–237

18 Magidovich, E., Orr, I., Fass, D., Abdu, U. and Yifrach, O. (2007) Intrinsic disorder in the C-terminal domain of the Shaker voltage-activated K^+ channel modulates its interaction with scaffold proteins. Proc. Natl Acad. Sci. U.S.A. **104**, 13022–13027

19 Vacic, V., Uversky, V. N., Dunker, A. K. and Lonardi, S. (2007) Composition profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics **8**, 211

20 Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. and Bax, A. (1995) NMR Pipe: a multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR **6**, 277–293

21 Permyakov, S. E., Millett, I. S., Doniach, S., Permyakov, E. A. and Uversky, V. N. (2003) Natively unfolded C-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin. Proteins **53**, 855–862

22 Csizmok, V., Bokor, M., Banki, P., Klement, E., Medzihradszky, K. F., Friedrich, P., Tompa, K. and Tompa, P. (2005) Primary contact sites in intrinsically unstructured proteins: the case of calpastatin and microtubule-associated protein 2. Biochemistry **44**, 3955–3964

23 Uversky, V. N. (2002) What does it mean to be natively unfolded? Eur. J. Biochem. **269**, 2–12

24 Receveur-Brechot, V., Bourhis, J. M., Uversky, V. N., Canard, B. and Longhi, S. (2006) Assessing protein disorder and induced folding. Proteins **62**, 24–45

25 Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W. et al. (2001) Intrinsically disordered protein. J. Mol. Graph. Model. **19**, 26–59

26 Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N. and Dunker, A. K. (2005) Comparing and combining predictors of mostly disordered proteins. Biochemistry **44**, 1989–2000

27 Uversky, V. N., Gillespie, J. R. and Fink, A. L. (2000) Why are ‘natively unfolded’ proteins unstructured under physiologic conditions? Proteins **41**, 415–427

28 Dyson, H. J. and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins. Curr. Opin. Struct. Biol. **12**, 54–60

29 Dyson, H. J. and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. **6**, 197–208

30 Dobson, C. M. (2003) Protein folding and misfolding. Nature **426**, 884–890

31 Bukau, B., Weissman, J. and Horwich, A. (2006) Molecular chaperones and protein quality control. Cell **125**, 443–451

32 Hartl, F. U., Bracher, A. and Hayer-Hartl, M. (2011) Molecular chaperones in protein folding and proteostasis. Nature **475**, 324–332

33 Calderone, C. T. and Williams, D. H. (2001) An enthalpic component in cooperativity: the relationship between enthalpy, entropy, and noncovalent structure in weak associations. J. Am. Chem. Soc. **123**, 6262–6267

34 He, X., Chow, D., Martick, M. M. and Garcia, K. C. (2001) Allosteric activation of a spring-loaded natriuretic peptide receptor dimer by hormone. Science **293**, 1657–1662

35 Tsvetkov, P., Reuven, N. and Shaul, Y. (2009) The nanny model for IDPs. Nat. Chem. Biol. **5**, 778–781

36 Uversky, V. N., Oldfield, C. J. and Dunker, A. K. (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J. Mol. Recognit. **18**, 343–384

37 Wright, P. E. and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. **293**, 321–331

38 Fuxreiter, M., Tompa, P., Simon, I., Uversky, V. N., Hansen, J. C. and Asturias, F. J. (2008) Malleable machines take shape in eukaryotic transcriptional regulation. Nat. Chem. Biol. **4**, 728–737

39 Gsponer, J., Futschik, M. E., Teichmann, S. A. and Babu, M. M. (2008) Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science **322**, 1365–1368

40 Shoemaker, B. A., Portman, J. J. and Wolynes, P. G. (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proc. Natl Acad. Sci. U.S.A. **97**, 8868–8873

Received 16 May 2012/28 May 2012; accepted 6 June 2012
Published as Immediate Publication 11 June 2012, doi 10.1042/BSR20120049