the FEBS Journal

Structural basis for protein *trans*-splicing by a bacterial intein-like domain – protein ligation without nucleophilic side chains

A. Sesilja Aranko*, Jesper S. Oeemig* and Hideo Iwaï

Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki, Finland

Keywords
BIL domain; Hint domain; NMR spectroscopy; protein ligation; protein *trans*-splicing

Correspondence
H. Iwaï, Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki, PO Box 65, Helsinki, FIN-00014, Finland
Fax: +358 9 191 59541
Tel: +358 9 191 59752
E-mail: hideo.iwai@helsinki.fi

*These authors contributed equally to this work

(Received 15 March 2013, revised 17 April 2013, accepted 22 April 2013)

doi:10.1111/febs.12307

Protein splicing in *trans* by split inteins has become a useful tool for protein engineering *in vivo* and *in vitro*. Inteins require Cys, Ser or Thr at the first residue of the C-terminal flanking sequence because a thiol or hydroxyl group in the side chains is a nucleophile indispensable for the *trans*-esterification step during protein splicing. Newly-identified distinct sequences with homology to the hedgehog/intein superfamily, called bacterial intein-like (BIL) domains, often do not have Cys, Ser, or Thr as the obligatory nucleophilic residue found in inteins. We demonstrated that BIL domains from *Clostridium thermocellum* (*Cth*) are proficient at protein splicing without any sequence changes. We determined the first solution NMR structure of a BIL domain, *Cth*BIL4, to guide engineering of split BIL domains for protein ligation. The newly-engineered split BIL domain could catalyze protein ligation by *trans*-splicing. Protein ligation without any nucleophilic residues of Cys, Ser and Thr could alleviate junction sequence requirements for protein *trans*-splicing imposed by split inteins and could broaden applications of protein ligation by protein *trans*-splicing.

Database
The resonance assignments and structure coordinates have been deposited in BMRB (18653) and RCSB (2LWY)

Introduction

Protein splicing is an intriguing post-translational modification of proteins in which an intervening sequence excises itself from the host protein, simultaneously joining the two divided host protein fragments [1–3]. Protein splicing and auto-processing post-translational modification of hedgehog proteins are often considered to be evolutionally related because both reactions are catalyzed by the hedgehog/intein (Hint) fold [4,5]. Although protein-splicing domains termed inteins catalyze self-excision from the host proteins and ligation of the two flanking sequences with a peptide bond, the C-terminal domains of hedgehog proteins catalyze the release of the N-terminal domain and lipid modification at the C-terminus of the N-terminal domain via an N-S acyl shift (Fig. 1) [5]. The hedgehog domain utilizes the same first-step N-S acyl shift of protein splicing but the C-terminal *trans*-esterification by the peptide chain in protein splicing is replaced by nucleophilic attack of a hydroxyl group of cholesterol [4]. Structural analysis has revealed the similarity of the three-dimensional architectures of the two domains with a horseshoe-like disk shape.

Abbreviations
Cth, *Clostridium thermocellum*; BIL, bacterial intein-like; GB1, the B1 domain of IgG binding protein G; Hh, hedgehog protein; Hint, hedgehog/intein; IMAC, ion-metal immobilized chromatography; IPTG, isopropyl thio-β-D-galactoside; *Npu*, *Nostoc punctiforme*. PDB, Protein Data Bank; PTS, protein *trans*-splicing; *Ssp*, *Synechocystis* sp. strain PCC6803; YFP, yellow fluorescent protein.

Rapid advances in genomic DNA sequencing have identified an increasing number of inteins in various organisms [6]. Database searches have also identified protein sequences with homology to the Hint domain in diverse bacterial species, which is distinct both from inteins and hedgehog proteins [7,8]. The newly-identified Hint domain has been termed bacterial intein-like (BIL) domain, which typically consists of 130–155 residues [7]. BIL domains can catalyze the cleavage of their host proteins, as well as protein splicing, indicating that BIL domains are closely related to inteins [7,9]. Although inteins are usually inserted into an active site of essential proteins that are often involved in DNA metabolism such as DNA polymerases, BIL domains are found in non-essential and secreted proteins. The biological functions of BIL domains are still obscure, although cleavage of the polypeptide chains appears to be an important function of BIL domains for the host proteins [9]. BIL domains have been categorized into three groups (i.e. A-, B- and C-types) based on distinct characteristic sequence motives [7,9,10]. Importantly, BIL domains often lack Cys, Ser or Thr at the +1 position, which corresponds to the first residue after the intein [i.e. the first residue of C-terminal flanking sequence (C-extein)]. These nucleophilic residues at the +1 position are indispensable for the *trans*-esterification step in protein splicing by inteins [3]. Despite the lack of a thiol or hydroxyl group in the side chain of the +1 residue, A-type BIL domains were demonstrated to catalyze protein splicing [7,9]. The efficiency of protein splicing catalyzed by BIL domains was found to be as low as 10–25% in model systems [7,9]. The absence of Cys, Ser and Thr at the +1 position makes BIL domains attractive for the development of protein ligation tools because protein ligation technology using split inteins that catalyze protein *trans*-splicing (PTS) is often restricted by the prerequisite of Cys, Ser or Thr at the ligation junction (Figs 1 and 2) [3,11]. Protein ligation by PTS without Cys, Ser and Thr could broaden the applications of PTS because it could significantly alleviate the sequence requirements for PTS [11–15]. Not only protein ligation by PTS, but also other *in vitro* ligation approaches, such as native chemical ligation, require an N-terminal cysteine in the C-terminal fragment as a nucleophile [16]. The limitations imposed by the ligation junction sequences also restrict the broad applications of other enzymatic approaches, including sortase-mediated ligation and subtiligase [17,18]. BIL domains identified in the genome sequences contain diverse amino acid types at the +1 position that cannot usually serve as a nucleophile like Cys, Ser or Thr [7]. Therefore, BIL domains catalyzing protein splicing without Cys/Ser/Thr at the +1 position are of potential importance for developing Cys/Ser/Thr-free protein ligation by *trans*-splicing using split BIL domains. Previously, 10 genes containing BIL domains have been identified in the genome sequence of *Clostridium*

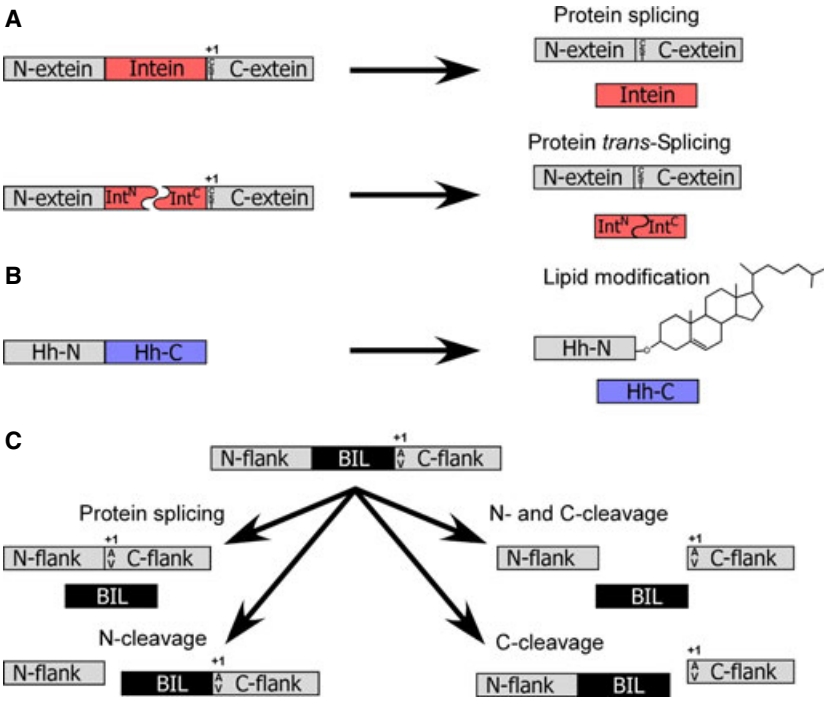

Fig. 1. Schematic drawing of the reactions catalyzed by (A) inteins and split inteins, (B) hedgehog protein (Hh) and (C) bacterial intein-like (BIL) domains.

Fig. 2. A sequence alignment of BIL domains from *C. thermocellum*. The secondary structures of *Cth*BIL4 obtained from the NMR structure are indicated on top of the alignment. Conserved intein sequence regions of blocks A, B, F and G are also aligned based on the structures. The residues are highlighted in accordance with the sequence identity among BIL domains (< 70% in dark grey, > 70% in grey, > 85% in light grey and 100% in white, respectively). h, hydrophobic residues (G, V, L, I, A,M); a, acidic residues (D, E); r, aromatic residues (F, Y, W); p, polar residues (S, T, C) (6). The two split sites are indicated by filled triangles (labelled with C42 and C16).

thermocellum (*Cth*) [7]. One of them, *Cth*BIL4, has been biochemically characterized, catalyzing predominantly N- and C-cleavage [9]. Despite the fact that *Cth*BILs lack a typical Cys, Ser or Thr at the +1 position found in inteins (Fig. 1), a minute amount of the spliced product was also observed with *Cth*BIL4 [9]. BIL domains in *C. thermocellum* are highly homologous, with sequence identities of >80% (Fig. 2). These BIL domains are found in proteins of unknown function (DUF1577).

To better understand the function of BIL domains and the splicing mechanism at atomic resolution, we determined the solution structure of an A-type BIL domain, *Cth*BIL4, by NMR spectroscopy. The first three-dimensional structure of a BIL domain suggests a strong relation to inteins and provides the structural basis for creating a new protein engineering tool.

Results

The NMR solution structure of *Cth*BIL4

For the structural analysis, we introduced a mutation of C1A in *Cth*BIL4 to prevent any possible cleavage and splicing reactions. A sequence of SMK was also included at the N-terminus of *Cth*BIL4 as an N-flanking sequence. The nearly complete resonances were assigned for the structure calculations (Fig. S1). The 20 NMR solution structures have been well defined with the backbone rmsd value of 0.34 ± 0.05 Å, which do not have any significantly disordered regions (Figs 3 and S2 and Table 1). *Cth*BIL4 consists of 12 β-strands without any helical region, folded into a horseshoe-like fold, in which the N- and C-termini are located at the centre of the structure as commonly observed in the Hint superfamily (Fig. 3a).

Structural comparison with inteins

The length of inteins varies significantly from 134 residues to > 1000 residues [6]. Inteins are typically 400 residues long as a result of an endonuclease domain insertion in the loop between blocks B and F [19]. The endonuclease domains can play an important role for intein propagation. Inteins are often considered to be functionless, selfish genetic elements because their host proteins are usually essential for the organisms and due to the presence of the endonuclease domains for propagation [20]. Inteins lacking an endonuclease domain are often called mini-inteins [21]. The lengths of the mini-inteins also vary from 134 to 200–300 residues. By contrast, BIL domains appear to be relatively invariable because BIL domains have typically between 133 and 155 residues. To identify the most structurally

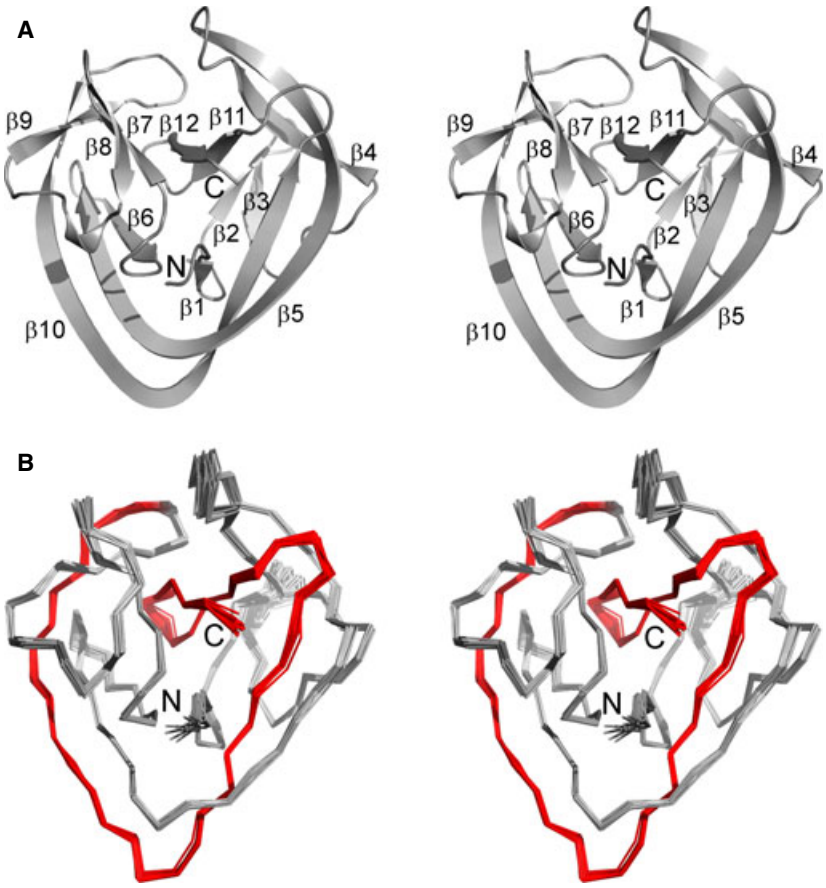

Fig. 3. The NMR solution structure of *Cth*BIL4 domain. (A) A stereo view of ribbon presentation of the NMR structure labelled with the secondary structures. (B) A stereo view of the 20 superimposed NMR structures. N- and C-termini are indicated. The C-terminal region corresponding to the split C42 fragment is coloured in red (see text).

similar structure, the NMR structure of *Cth*BIL4 was subjected to a DALI server search [22]. The DALI server identified the structure of DnaB intein from *Synechocystis* sp. PCC6803 (*Ssp*) [Protein Data Bank (PDB): 1MI8] with the highest *Z*-score of 19.2, covering 131 residues with an rmsd of 1.6 Å (Fig. 4a) [21]. PI-*Mtu*I (RecA intein) and PI-*Pko*II (PDB: 3IFJ, 2CW7) also have relatively high *Z*-scores [23,24]. This suggests that the backbone conformation of *Cth*BIL4 is similar to inteins, even though the BIL domain lacks an endonuclease domain. Not only the backbone conformations, but also all the active site residues can be well superimposed between *Cth*BIL4 and *Ssp*DnaB intein (Fig. 4c). The notable difference with inteins is Asn115 in *Cth*BIL4, located at the position of the highly-conserved Asp at block F among inteins, which plays a critical role in protein splicing in some inteins (Figs 2 and 4) [21,25,26].

Comparison with hedgehog protein

Another member of Hint superfamily is the C-terminal domain of hedgehog protein (Hh-C), which catalyzes lipid modification of the N-terminal signalling domain of hedgehog protein (Hh-N) (Fig. 1b). The DALI server gave a *Z*-score of 16.7 for Hh-C (PDB: 1AT0), covering 128 residues with an rmsd of 2.0 Å [4]. Structure alignment between *Cth*BIL4 and Hh-C revealed that most of the variations are located in the loop regions (Fig. 4b). The largest differences of the regular secondary structures are β5 and β10 in *Cth*BIL4, which correspond to β4a and β4b in Hh-C, respectively [4]. The structural deviations are caused by the differences in the lengths of the β-strands. Although β5 and β10 in *Cth*BIL4 consist of nine residues (residues 45–53) and 12 residues (residues 102–113) in *Cth*BIL4, the corresponding β4a and β4b in the Hh-C structure consist of 11 residues (residues 302–312) and nine residues (residues 373–381), respectively. These differences in length are accommodated by slight shifts of the β-strands, increasing the rmsd. The superposition of the two structures indicates that the active site residues Cys1, Thr65 and His68 in *Cth*BIL4 correspond to Cys258, Thr326 and His329 in Hh-C, respectively (Fig. 4d). The highly-conserved Asp in block F of inteins is replaced with Asn115 in *Cth*BIL4 but with Ala383 in

Table 1. Experimental data for the NMR structure calculation and the structural statistics.

Quantity	Value
NOE upper distance limits	
Short range NOE ($i-j \leq 1$)	1662
Medium-range NOE ($1 < i-j < 5$)	465
Long-range NOE ($i-j \geq 5$)	1608
Residual CYANA target function	0.38 ± 0.03
Residual NOE violation	
Number ≥ 0.2 Å	2 ± 2
Maximum (Å)	0.35 ± 0.25
Amber energies (kcal·mol^{-1})	
Total	−91 931 ± 1417
van der Waals	14 982 ± 269
Electrostatic	−122 650 ± 1759
rmsd from ideal geometry	
Bond length (Å)	0.0238 ± 0.0001
Bond angles (°)	2.142 ± 0.017
rmsd to mean coordinate	
Backbone 1–135 (Å)	0.34 ± 0.05
Heavy atoms 1–135 (Å)	0.65 ± 0.06
Ramachandran plot statistics (%)[a]	
Most favoured regions	91.7
Additional allowed regions	8.2
Generously allowed regions	0.1
Disallowed regions	0.0

[a] Derived by PROCHECK-NMR [46].

Hh-C. Asn135 in *Cth*BIL4 corresponds to the last residue of intein, which cleaves the branched intermediate by Asn cyclization. This Asn residue is absent in Hh-C because it does not require any cleavage of C-extein. The structure of the active sites also clearly indicates the closer relationship of BIL domains to inteins rather than Hh-C.

Fully functional protein splicing activity of *Cth*BIL4

The predominant function of BIL domains was considered to be cleavage of the host proteins. Because some type-A BIL domains possess intein signature motifs except for the +1 residue, protein-splicing activity could be partly restored when Ala at the +1 position was replaced with Cys in the BIL domain from *Magnetospirillum magnetotacticum* [27]. The superposition of *Cth*BIL4 structure and intein structures allows us to compare the characteristic intein signatures found in blocks B and F with atomic precision (Fig. 2). Structural comparison also confirmed that the position of Asp in the block F found in many inteins corresponds to Asn115 in *Cth*BIL4 (Fig. 4c). The Asp in the block F was considered to be very important for the splicing reaction because replacement by Asn resulted in predominant cleavage reactions, suggesting that this residue might be important for the cleavage of *Cth*BIL4 [21,25]. Interestingly, when a mutation of A+1C at the +1 splicing junction was introduced in our model system that uses the B1 domain of immunoglobulin IgG binding protein G (GB1) as N- and C-flanking sequences, *Cth*BIL4 spliced with an efficiency of > 90% (defined as the ratio between His-tagged ligated product and the remaining precursor and cleaved products with His-tag), dominantly producing the *cis*-spliced product of H_6-GB1-GB1 (Fig. S3). It also had negligible N- and C-cleavages (Fig. 5a). *Cth*BIL4 was thus fully converted to an 'intein' with efficient *cis*-splicing activity. *Cth*BIL4 is fully proficient at protein splicing without any sequence changes within *Cth*BIL4, although the following +1 residue is essential for splicing. This suggests that Asp in the block F found in inteins is not required for splicing by *Cth*BIL4. We also tested another BIL domain, BIL5 from *C. thermocellum* (*Cth*BIL5), for which the sequence identity to *Cth*BIL4 is 89% (Fig. 2). *Cth*BIL5 has Val at the +1 position instead of Ala in *Cth*BIL4. The effect of the mutation from Val to Cys (V+1C) was similar to that of *Cth*BIL4, producing predominantly the *cis*-splicing product (Fig. 5b). BIL domains of *C. thermocellum* are capable of protein splicing depending on the following +1 position despite the Asn in block F. We also tested the effect of a mutation from Asn115 to Asp115 in *Cth*BIL4 with +1A, resulting in predominant C-cleavage with no *cis*-splicing product (Fig. 5c). The functional role of Asn in the block F in *Cth*BIL appears to be different from Asp in block F, which is highly conserved among many inteins despite the similar three-dimensional coordination [25]. Considering that many inteins do not necessarily have efficient splicing activity even with native junction sequences, and also induce considerable side-reactions of N- and/or C-cleavage [28], *Cth*BILs are fully functional protein splicing domains similar to inteins, depending on the +1 residue despite their distinct sequence homology from inteins [7].

A possible function of *Cth*BIL4

BIL domains lack endonuclease domains, unlike most inteins. The endonuclease domain in inteins presumably plays an important role for the spread of inteins by horizontal transfer [20]. BIL domains might have evolved from a mini-intein and are likely to have been fixed for some function during evolution. One of the proposed functions of BIL domains is to increase the variability of host proteins [9]. *Cth*BIL4 and *Cth*BIL5 act as functional intein-like protein splicing domains

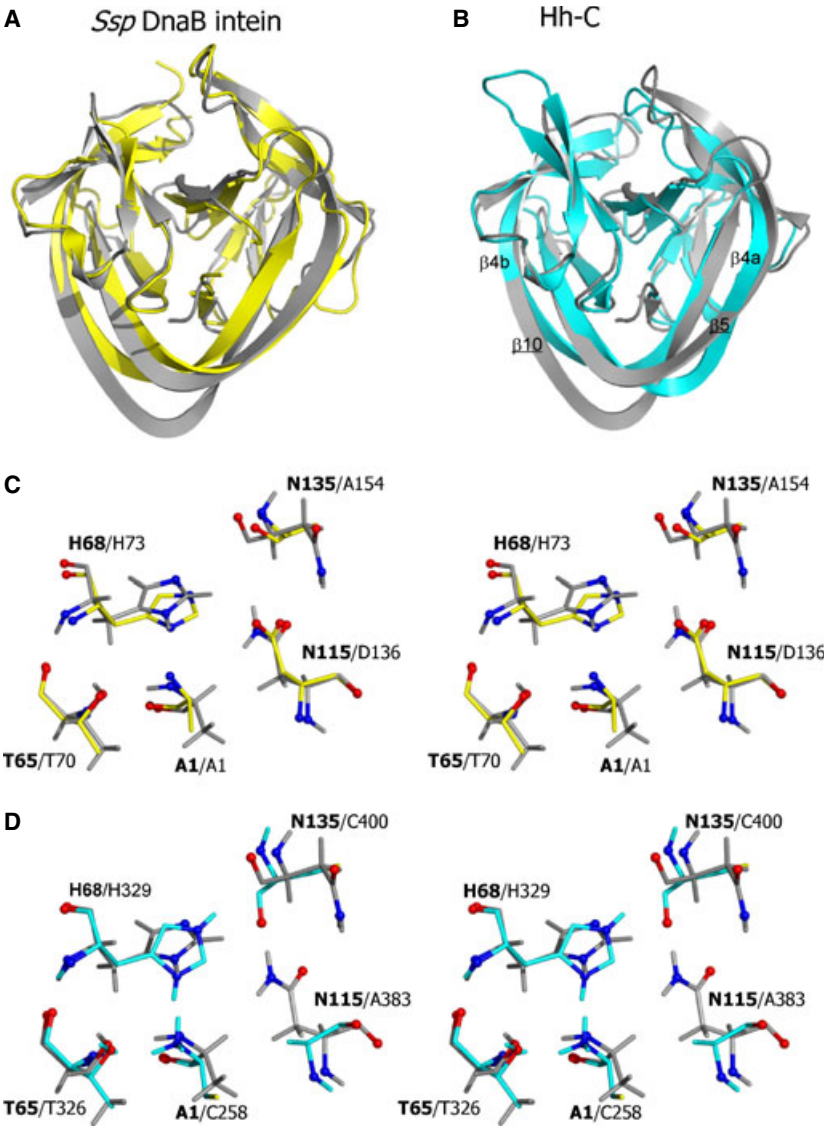

Fig. 4. Comparisons of *Cth*BIL4 structure and other Hint folds. (A) Superposition of the two structures of *Cth*BIL4 (grey) and *Ssp*DnaB intein (PDB: 1MI8) (yellow). (B) Superposition of the two structures of *Cth*BIL4 (grey) and hedgehog C-terminal domain (PDB: 1AT0) (cyan). (C) Comparison of the active sites of *Cth*BIL4 and *Ssp* DnaB intein or (D) *Cth*BIL4 and Hh-C in stereo view. Active site residues from *Cth*BIL4 (bold), *Ssp* DnaB intein, and Hh-C are labelled.

when Cys at the +1 position is introduced. We further tested the mutations of other nucleophilic residues of Ser and Thr at the +1 position, as also used for protein splicing by many inteins. However, these mutations did not increase protein splicing but diminished the C-cleavage product (H_6-GB1-BIL) (Fig. 5a). We also tested Gly at the +1 residue, showing no improvement of protein splicing. These mutations suggest that the +1 position with *Cth*BIL4 might be optimized to have all N-, C-cleavage and protein splicing reactions, thereby increasing the molecular repertories of the host protein, as suggested previously [9]. Because the host protein is a protein of unknown function, the biological function of BIL domains in *C. thermocellum* remains to be revealed. However, our experiments demonstrate that *Cth*BIL4 is fully capable of protein splicing, although the protein splicing function is not fully exploited by retaining Ala at the +1 position. We speculate that this might be because the variety of different cleaved and spliced molecules is important for host protein function. Therefore, the relative proportions of the different molecules produced by BIL domains are tuned by the +1 position.

Protein ligation by *Cth*BIL4

Cis-splicing of *Cth*BIL4 and *Cth*BIL5 bearing Cys at the +1 position is efficient with very little cleavage and is comparable to highly efficient inteins [28]. This observation prompted us to create split BIL domains for *trans*-splicing. We split *Cth*BIL4 into two halves based on the structural information from the solution

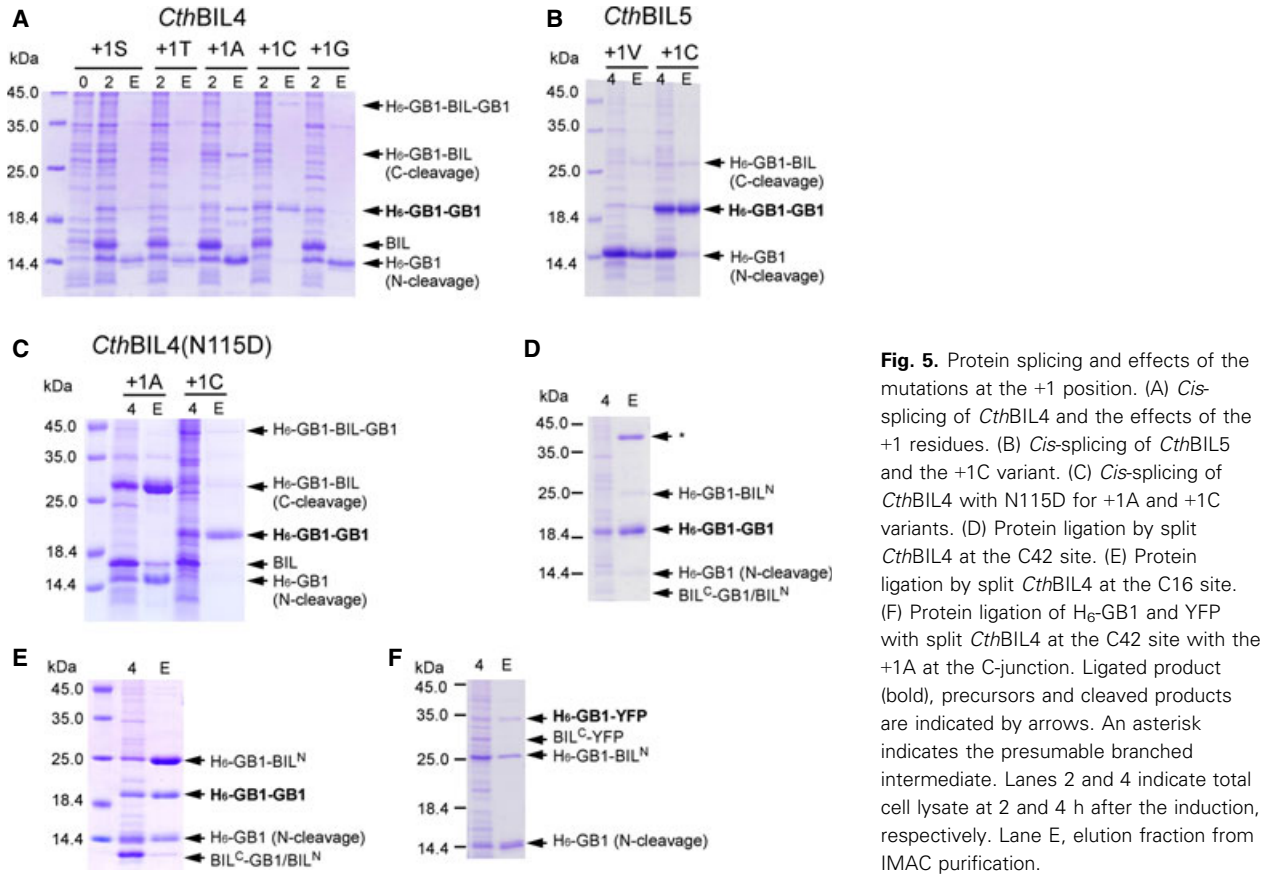

Fig. 5. Protein splicing and effects of the mutations at the +1 position. (A) *Cis*-splicing of *Cth*BIL4 and the effects of the +1 residues. (B) *Cis*-splicing of *Cth*BIL5 and the +1C variant. (C) *Cis*-splicing of *Cth*BIL4 with N115D for +1A and +1C variants. (D) Protein ligation by split *Cth*BIL4 at the C42 site. (E) Protein ligation by split *Cth*BIL4 at the C16 site. (F) Protein ligation of H_6-GB1 and YFP with split *Cth*BIL4 at the C42 site with the +1A at the C-junction. Ligated product (bold), precursors and cleaved products are indicated by arrows. An asterisk indicates the presumable branched intermediate. Lanes 2 and 4 indicate total cell lysate at 2 and 4 h after the induction, respectively. Lane E, elution fraction from IMAC purification.

structure of *Cth*BIL4, as previously demonstrated with an intein [29]. The first split site in *Cth*BIL4 was introduced at the front of β10 (termed C42) (Fig. 3), which corresponds structurally to the naturally split site of the *Nostoc punctiforme* (*Npu*) DnaE intein [29,30]. This new split BIL was capable of *trans*-splicing with good yields using a model system of GB1s, producing *trans*-spliced product of H_6-GB1-GB1 by *in vivo* ligation (Fig. 5d). We also shortened the C-terminal fragment further to 16 residues (termed C16) by splitting after residue 119 between β10 and β11 of *Cth*BIL4 because this region also indicated internal flexibility (Fig. S2). This split site corresponds to an engineered split *Npu* DnaE intein [31]. The newly-split BIL with the C16 site could also splice in *trans in vivo*, although there were more cleaved products than with the C42 site (Fig. 5e). For *in vitro* protein ligation by BIL-mediated *trans*-splicing, the C42 site was inappropriate because split fragments became insoluble. However, split fragments at the C16 site can be purified under nondenaturing conditions and might be used for the *in vitro* ligation reaction despite its lower yield and slow reaction (Fig. S4a). These split BIL domains could be used for *in vivo* and *in vitro* protein ligation by PTS.

Cys/Ser/Thr-free protein ligation by *Cth*BIL4

Although inteins invariably require Cys, Ser or Thr at the +1 position of C-extein for protein splicing [32], BIL domains contain various amino acid types at the +1 position of the C-flanking sequence. Albeit with poor efficiency (10–25%), BIL domains have been demonstrated to splice with Ala or Val at the +1 position [7,9]. PTS by split BILs without any Cys, Ser and Thr at the +1 position could significantly alleviate the sequence requirements for protein ligation by PTS with inteins even when the yield might be low. Therefore, we tested whether split BIL domains can splice in *trans* without Cys at the +1 position by using *Cth*BIL4 split at the C42 site. We used His-tagged GB1 and yellow fluorescent protein (YFP) as N- and C-flanking sequences, respectively. The N-terminal split BIL (BIL^N) was fused to an N-terminally His-tagged GB1, which is encoded in a plasmid with kanamycin resistance. The C-terminal split BIL (BIL^C) was fused with YFP in the other compatible plasmid with ampicillin resistance. Both precursor proteins were overexpressed simultaneously to test PTS between them in *Escherichia coli* [31]. His-tagged proteins were purified

by immobilized metal ion affinity chromatography (IMAC) and analyzed by SDS-PAGE (Fig. 5f). The purified products by IMAC included the ligated product of H_6-GB1-YFP (Fig. 5f). The ligation was also confirmed by MALDI-MS and by monoclonal anti-His-tag and anti-green fluorescent protein antibodies (Fig. S5). We obtained approximately 0.5 $mg \cdot L^{-1}$ of the ligated H_6-GB1-YFP. Split BIL at the C16 site, which is less productive but more soluble than the C42 site, was also used for *in vitro* ligation without +1Cys, producing a minute amount of the spliced product (Fig. S4b).

Towards better protein ligation by BIL domains

Protein ligation efficiency without +1Cys is very low for most applications. To improve the efficiency by rational protein engineering, it is necessary to understand the mechanism at atomic resolution. The mechanism of protein splicing by A-type BIL domains has been proposed previously [9]. The first step is the N-S acyl shift reaction common for Hint domains. The second step is presumably C-terminal cleavage upon Asn cyclization (Fig. 6a). The C-terminally cleaved products could be diffused away, resulting in C-cleaved product. The N-terminal amino group released by C-cleavage could also attack the thioester group for aminolysis, producing the ligation product (splicing product). Aminolysis of the thioester intermediate of canonical inteins with +1Ala has been investigated extensively [33]. This aminolysis has been also used for the amidation of recombinant proteins because this thioester group could be attacked by any amine in the vicinity [33,34]. In the case of wild-type *Cth*BIL4 with a sequence of 'AN' at the C-terminal junction, the carboxamide group of Asn at the +2 position (the second residue after the BIL domain), which is in very close proximity, could also attack the same thioester group, forming an imide group (Fig. 6a, pathway 1). The imide group can be easily hydrolyzed at a higher pH. The presumable 'ligation' band observed with the C-terminal 'AN' junction sequence diminished during storage in buffer at pH 8 with incubation at room temperature or in SDS loading buffer at −20 °C (Fig. 6b). This is probably a result of hydrolysis of the imide group. This product was stable only under acidic conditions, supporting the imide formation (Fig. 6c). When Asn at the +2 position was replaced by Ser, the 'ligation' band did not diminish, thereby suggesting that stable amide bond formation by aminolysis with the N-terminal amino group, as confirmed previously [9]. We consider that, in the case of +1Cys, the thiol group of cysteine residue could directly attack the thioester group after N-S acyl shift for *trans*-esterification similar to inteins, which might occur even earlier than Asn cyclization because we observed a significant reduction of C-cleaved product and the N115D mutation did not affect *cis*-splicing with +1Cys (Figs 5c and 6b). For better protein ligation without Cys/Ser/Thr at the +1 position, BIL domains might be improved by removing amino group in the vicinity and preventing the diffusion of the C-cleaved product before aminolysis.

Discussion

The protein splicing mechanism catalyzed by inteins has several variations for achieving protein splicing because several diversified mechanisms have been identified and classified into three classes [35]. Even among the inteins classified as class I inteins, the residues in the vicinity of the catalytic site are not strictly conserved [6,36]. Protein splicing can thus be accomplished by different residues in individual inteins. BIL domains were originally identified as homologous sequences that are distinct from characteristic intein sequences. Particularly, indispensable Cys, Ser or Thr at the +1 position for protein splicing by inteins is absent among many BIL domains. However, our experiments demonstrated that *Cth*BIL4 and *Cth*BIL5 are fully capable of intein-like protein splicing activity, although missing Cys at the +1 position for protein splicing. The *Cth*BIL domain is yet another sequence variant that can perform protein splicing efficiently. This suggests that the roles of the residues in the vicinity of the splicing site vary considerably not only among inteins, but also among the Hint superfamily. The functional roles of these residues in the vicinity of the splicing site appear to be very specific to individual inteins/BIL domains, although the key steps in canonical protein splicing mechanism (e.g. the N-S acyl shift step) are probably shared with many inteins and intein-related proteins but performed in a slightly different manner [32,36]. There might be undiscovered intein-related sequences that could also catalyze protein splicing with or without the typical Cys/Ser/Thr splicing motif at the +1 position. Importantly, we demonstrated that protein ligation could be achieved by using a split BIL domain without any Cys, Ser and Thr in the C-terminal flanking sequence, albeit with a very low efficiency. This result opens a new possibility for PTS catalyzed by BIL domains. The optimally engineered BIL domains might be able to overcome the limitation imposed by the splicing junction requirements of intein-mediated PTS [11–15]. A better understanding

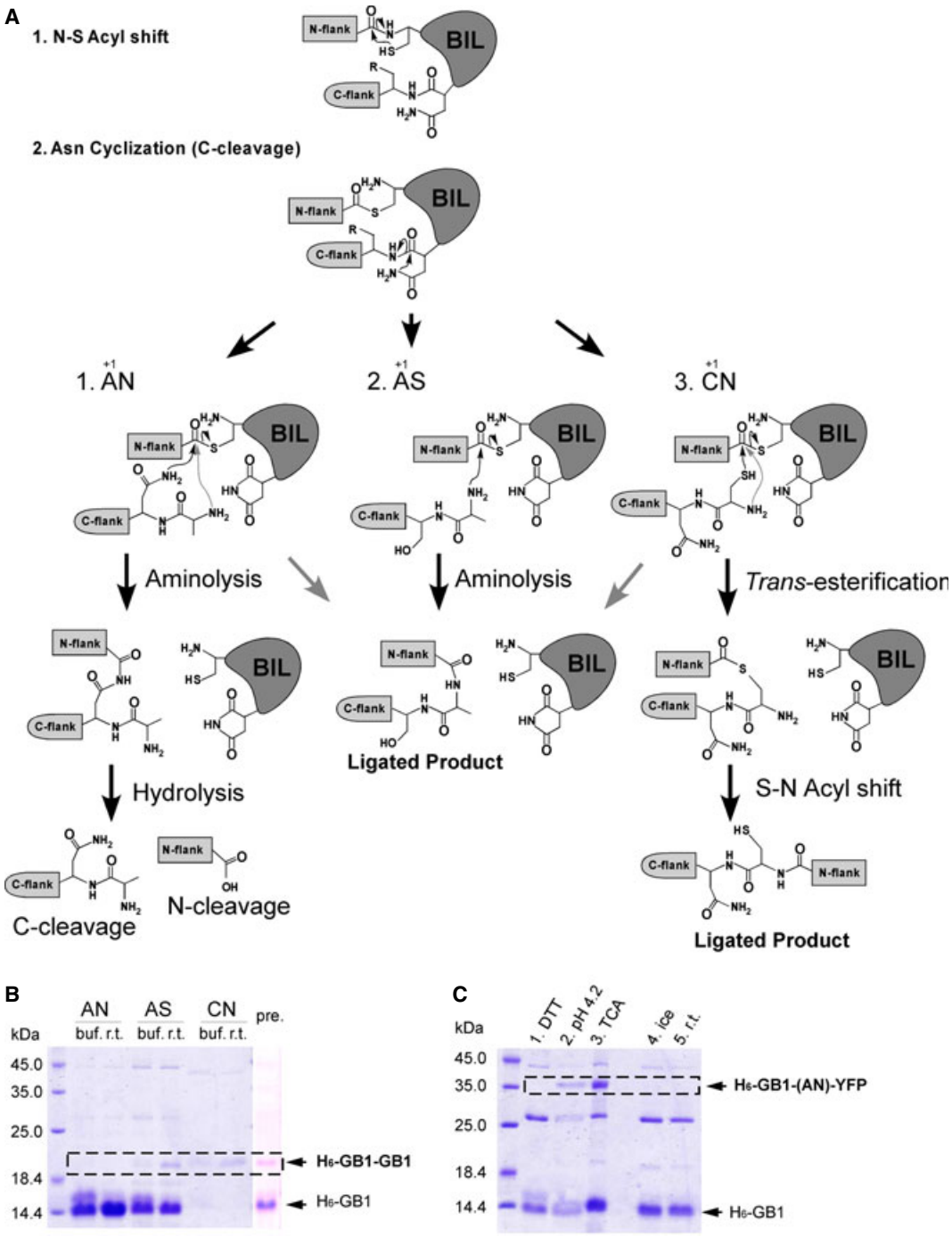

Fig. 6. (A) Proposed mechanisms with three different junction sequences (AN, AS and CN). Grey arrows indicate the possible alternative reaction pathways of aminolysis by N-terminal amine, which could also be present. (B) Stability of the 'ligated' bands with different junction sequences. Lanes buf., samples stored at −20 °C in SDS loading buffer overnight; r.t., samples incubated at room temperature overnight in elution buffer (pH 8.0); pre, pre-incubation sample immediately after IMAC purification. (C) Stability of the 'ligated' band of H_6-GB1-YFP with the 'AN' junction sequence under various conditions. Lane 1, elution from IMAC with 1 mM dithithreitol; lane 2, 1 : 1 mixture of elution from IMAC with 0.5 M sodium phosphate (pH 4.2); lane 3, precipitated sample by 50% trichloroacetic acid; lane 4, incubated on ice overnight in elution buffer (50 mM sodium phosphate, 300 mM NaCl, 250 mM imidazole, pH 8); lane 5, incubated at room temperature overnight in elution buffer.

of the protein splicing mechanism by BIL domains and rational protein engineering could further improve the ligation efficiency and enable the practical application of split BIL domains.

Materials and methods

Construction of plasmids

BIL4 gene (NCBI GI number: 23020817) was amplified from genomic DNA of *C. thermocellum* (DSM-1237) using PCR with the oligonucleotides: I53: 5′-A GGA TCC ATG AAG TGC TTT GTT GCA GGC-3′ and I54: 5′-TA GGT ACC ATA ATT TGC ATT ATG CAC CAA TAC-3′ (for pSCFDuet7) or I206: 5′-TTG GAT CCT GCT TTG TTG CAG GCA CG-3′ and I207: 5′-GA GGT ACC ACG AGA TGC ATT ATG CAC CAA TAC-3′ (for pSADuet712). The gene of *Cth*BIL5 was also amplified from the genomic DNA by PCR with the oligonucleotides: I214: 5′-GA GGA TCC AAG TGC TTT GTT GCA GGC ACG ATG-3′ and I215: 5′-CG GGT ACC ATA TTT TAC ATT ATG AAC CAA AAC-3′.

The PCR products were digested with restriction enzymes of *Bam*HI and *Kpn*I and ligated into pSKDuet16, resulting pSADuet712 or pSCFDuet7 (*Cth*BIL4) and pSCFDuet98 (*Cth*BIL5) [28]. The constructs for testing the *cis*-splicing contain an N-terminally hexahistidine-tagged GB1 (H_6-GB1) as the N-flanking sequence and a GB1 as the C-flanking sequence.

Mutagenesis of the +1 position with *Cth*BIL4 and *Cth*BIL5

For construction of the plasmids encoding the +1 variants of *Cth*BIL4, the oligonucleotides used for the PCR were: for +1C variant (pSADuet714), I53 and I219: 5′-TA GGT ACC ATA ATT ACA ATT ATG CAC CAA TAC TTC-3′; for +1G variant (pSADuet715), I53 and I220: 5′-TA GGT ACC ATA ATT ACC ATT ATG CAC CAA TAC TTC-3′; for +1S variant (pJODuet50), I53 and I234: 5′-TA GGT ACC ATA ATT AGA ATT ATG CAC CAA TAC TTC-3′; for +1T variant (pJODuet51), I53 and I235: 5′-TA GGT ACC ATA ATT AGT ATT ATG CAC CAA TAC TTC-3′. The PCR products were digested with *Bam*HI and *Kpn*I and ligated into pSKDuet16 using the same restriction sites. The +1C variant for *Cth*BIL5 was constructed with the same procedure using the oligonucleotides I214 and I243, 5′-CG GGT ACC ATA TTT ACA ATT ATG AAC CAA AAC-3′, resulting in pSADuet733.

N115D mutation

Plasmid pJODuet49 harbouring H_6-GB1-*Cth*BIL4-GB1 with the N115D mutation and +1Ala was constructed from pSCFDuet7 by inverse PCR using oligonucleotides I223: 5′-CCT GTT AAA GTT TAT GAT TTT AAA GTA GAT G-3′ and I224: 5′-CAT CTA CTT TAA AAT CAT AAA CTT TAA CAG G-3′. For expression of H_6-GB1-*Cth*BIL4-GB1 with N115D mutation and +1Cys, pJODuet52 was constructed using the two oligonucleotides I53 and I219 from pJODuet49.

Split *Cth*BIL4 for *trans*-splicing

Split *Cth*BIL4 was constructed as follows. For the N-terminal split fragment, the gene of N-terminal split BIL4 was amplified by PCR from pSADuet712 with the oligonucleotides, DuetMCS1-fw: 5′-GGA TCT CGA CGC TCT CCC T-3′ and I277: 5′-TAC AAG CTT ATC TTG AAT CAA GCA G-’ (for the C42 site), or DuetMCS1-fw and I244: 5′-TCC AAG CTT AAT CTA CTT TAA AAT TAT AAA C-3′ (for the C16 site), then cloned into pHYRSF-1 using *Bam*HI and *Hind*III sites, resulting in pSARSF740 (C42 site) and pSARSF725 (C16 site) [37]. The gene of the C-terminal split *Cth*BIL4 was also amplified from pSCFDuet7, together with GB1 by PCR using the oligonucleotides, I276: 5′-TT CAT ATG GGC AAT GTT TTA GTG-3′ and SZ015: 5′-TGC CAA GCT TAT TCC GTT ACG GTG-3′ for the C42 site, or I245: 5′-A CAT ATG GAC TTC CAT ACT TAT CA-3′ and SZ015 for the C16 site. The amplified genes were digested with *Nde*I and *Hind*III and ligated into pSKBAD2, resulting in pSABAD739 (C42) and pSABAD726 (C16) [37].

Protein splicing and cleavage analysis by IMAC purification

Protein splicing and cleavage by BIL domains were analyzed by expressing precursor proteins from the constructs described above. The proteins were expressed in 5-mL cultures from *E. coli* ER2566 cells harbouring the plasmid. The cells were grown and induced in log-phase with 1 mM isopropyl thio-β-D-galactoside (IPTG) for 2 or 4 h. The cells were harvested by centrifugation at 4500 ***g*** at 4 °C for 10 min for further purification. The cell pellets were suspended in 100 μL of B-PER® Bacterial Protein Extraction Reagent (Thermo Scientific, Waltham, MA, USA) and incubated at room temperature for 10 min. The cell lysate was then cleared by centrifugation at 15 000 ***g*** for 5 min. The cleared supernatant was loaded on a pre-equilibrated nickel-nitrilotriacetic acid spin column (Qiagen, Valencia, CA, USA). Unbound proteins were washed away with 50 mM sodium phosphate, 300 mM NaCl and 30 mM imidazole (pH 8.0). Bound proteins were eluded from the spin column using 150 μl of 50 mM sodium phosphate, 250 mM imidazole and 300 mM NaCl (pH 8.0). The elution fractions were diluted with SDS loading buffer containing 0.5 mM *tris*(2-carboxyethyl)phosphine and analyzed by 18% SDS/PAGE and stained with PhastGel™ Blue R (GE Healthcare, Milwaukee, WI, USA) for visualization.

Protein production of ^{13}C, ^{15}N labelled *Cth*BIL4

For the NMR sample preparation, the gene of *Cth*BIL4 was amplified using the oligonucleotides, I69: 5′-A GGA TCC ATG AAA GCC TTT GTT GCA GGC ACG ATG-3′ and I70: 5′-ACA AGC TTA ATT ATG CAC CAA TAC TTC ATT ATC-3′. The PCR product was cloned into pHYRSF53-36 vector between *Bam*HI and *Hind*III restriction sites for the fusion with Sumo domain tag [38]. The resulting plasmid of pSCFRSF17 encodes a fusion protein of the N-terminally H_6-tagged Sumo and *Cth*BIL4 bearing a C1A mutation, together with two-residue N-extein of MK and no C-extein. For the doubly ^{15}N-, ^{13}C-labelled sample, *E. coli* ER2566 strain harbouring pSCFRSF17 was grown in M9 medium containing 25 μg·mL^{-1} kanamycin, 0.8 g·L^{-1} $^{15}NH_4Cl$ as the sole nitrogen source, and 2 g·L^{-1} ^{13}C-D-glucose as the sole carbon source. The cells were grown at 37 °C until OD_{600} of 0.6 was reached, and subsequently induced with a final concentration of 0.5 mM IPTG. The protein was induced for another 4 h before harvesting by centrifugation at 6700 ***g*** at 4 °C for 10 min. The cell pellet was resuspended in 50 mM sodium phosphate and 300 mM NaCl buffer (pH 8.0) and flash-frozen in liquid nitrogen for storage at −75 °C. The cells were thawed and lysed by ultrasonication. The cell debris was cleared by centrifugation at 42 800 ***g*** at 4 °C for 50 min. The supernatant was applied on a 5-mL HisTrap FF column (GE Healthcare) after filtration with a 0.45-μm filter. The column was washed with 50 mM sodium phosphate and 300 mM NaCl (pH 8.0). Bound proteins were eluded from the column by applying a linear gradient of 50–250 mM imidazole in 50 mM sodium phosphate, 300 mM NaCl and 50 mM imidazole (pH 8.0). The eluted fractions containing the protein with the expected size were dialyzed against phosphate buffered saline at 4 °C. After dialysis overnight, N-terminally H_6-tagged Ubiquitin-like-specific protease 1 was added to the solution to cleave off H_6-Sumo purification tag [38]. The digested solution was loaded onto a 5-mL HisTrap FF column. The flow through fractions containing the labelled *Cth*BIL4 were collected and dialyzed against 20 mM sodium phosphate buffer (pH 6.0). The protein was concentrated with a centrifugal filtering device with a molecular weight cut-off of 3000 (Millipore, Billerica, MA, USA) and transferred into a microcell NMR tube (Shigemi. Inc., Allison Park, PA, USA).

NMR spectroscopy and resonance assignment

All NMR measurements were performed on INOVA 600 or 800 MHz spectrometers (Varian Inc., Palo Alto, CA, USA) equipped with cryogenic probeheads. The NMR experiments were recorded at 298 K with a 1.2 mM protein solution in 20 mM sodium phosphate (pH 6.0). For the resonance assignments, a set of two-dimensional and 3D spectra were recorded: [^1H,^{15}N]-HSQC, aliphatic CT-[^1H,^{13}C]-HSQC, HNCACB, CBCA(CO)NH, HNCO, HN(CA)CO, ^{15}N-edited TOCSY with a mixing time of 50 ms, HN(CO)HAHB, H(C)CH-COSY, H(C)CH-TOCSY with a mixing time of 50 ms, (HB)CB(CGCD)HD, (HB)CB(CGCDCE)HE, aromatic CT-[^1H,^{13}C]-HSQC.

For structure determination, a ^{15}N-edited NOESY-HSQC and aliphatic ^{13}C-edited NOESY-HSQC spectra, both with a mixing time of 80 ms, were recorded at ^1H frequency of 800 MHz. T_1 and T_2 relaxation and heteronuclear $^{15}N\{^1H\}$-NOEs measurements were performed at ^1H frequency of 600 MHz [39]. $T_1(^{15}N)$ relaxation rates were determined using the relaxation delays of 10, 20, 30, 40, 50, 70, 90, 110, 130, 150 and 210 ms. T_2 (^{15}N) relaxation rates were determined using a Carr–Purcell–Meiboom–Gill based pulse sequence with relaxation delays 10, 30, 50, 70, 90, 110, 150 and 210 ms. Heteronuclear $^{15}N\{^1H\}$-NOEs were determined by comparing peak intensities in HSQC spectra with and without ^1H saturation for 3.5 s.

All spectra were processed using NMRpipe [40] and resonance assignment was performed using CCPNMR, version 2.3 [41]. The backbone resonance assignments (H^N, N, HA, CA, CO, HB, and CB) of *Cth*BIL4 were completely assigned, except for Ser-3. For Met-2 and Lys-1, H^N and N resonance assignments are also missing. In total, 97% of all the expected resonances in *Cth*BIL4 were completed (Fig. S1). The NMR structure was calculated using CYANA, version 3.0, with an automatic NOE assignment procedure [42]. The chemical shift and unassigned NOE peak lists were used as sole inputs for the calculation [42]. In each calculation, 100 structures were calculated. The 20 structures with the lowest CYANA target function were chosen for energy refinement. The final structure was energy refined using AMBER in a 10 Å water shell [43]. The quality of the structures was validated using NMR-CING [44]. The statistics from the structure calculation are listed in Table 1.

In vitro ligation by split *Cth*BIL4 at the C16 site

C-terminal precursor proteins carrying split *Cth*BIL4 at the C16 site and GB1 as the flanking sequence were constructed in pHYRSF-1 from pSADuet714 (+1Ala) and pSADuet712 (+1Cys) by amplifying the DNA fragment with two oligonucleotides I245 and SZ015. The resulted plasmids carry H_6-*Cth*BIL4$_{C16}$-GB1 with +1Cys (pSARSF726) or +1Ala (pSARSF778).

Split precursor proteins were purified from *E. coli* cells carrying pSARSF725, pSARSF726, or pSARSF778. The cells were growth in 0.5 L of LB medium supplemented with 25 μg·mL^{-1} kanamycin until OD_{600} of 0.5-0.6 was reached, and then induced with a final concentration of 1 mM IPTG for 4 h at 37 °C. The cells were harvested by centrifuging for 10 min at 4690 ***g***. The cell pellets were resuspended with lysis buffer (300 mM KCl, 50 mM KH_2PO_4, 5 mM imidazole, pH 8.0) (pSARSF725 and pSARSF726) or into buffer A (50 mM NaPi, 300 mM NaCl, pH 8.0)

(pSARSF778). The precursor proteins were purified by IMAC after cell lysis by ultrasonification and removal of cell debris by centrifugation at 48 000 ***g*** for 50 min. Elution fractions from IMAC were dialyzed against ligation buffer (0.5 M NaCl, 10 mM Tris-HCl, pH 7.0, 1 mM EDTA) overnight at 4 °C.

Purified split precursor proteins were mixed at a final concentration of 80 μM with an equal molar ratio in ligation buffer with a final concentration of 0.5 mM Tris(2-carboxyethyl)phosphine hydrochloride. The *in vitro* reaction was followed at 37 °C by taking samples after mixing the two fragments. The samples were analyzed by 18% SDS/PAGE. The band intensities were quantified using IMAGEJ (NIH, Bethesda, MD, USA).

Split *Cth*BIL4 for protein ligation of GB1 and YFP

pSARSF373 was constructed from pHYDuet183 replacing the first three residues of 'GSS' by 'RGS' [28]. The gene of H_6-GB1-BIL4^N at the C42 site was amplified with oligonucleotides I277 and DuetMCS1-fw from pSADuet712 and then cloned into pSARSF373 with *Bam*HI and *Hind*III sites, resulting in pSARSF373-740. The gene of YFP (Venus) [45] was amplified from pHWV using the oligonucleotides SA001: 5′-GTG GTA CCG GCA AGG GCG AGG AGC-3′ and HK031: 5′-CGC AAG CTT AAG TGA TCC CGG CGG CGG-3′ and the gene of the C-terminal split *Cth*BIL4 was amplified using the oligonucleotides of I277 and I276. These genes were cloned into a pBAD vector, resulting in pSABAD128-771. The two compatible plasmids of pSABAD128-771 and pSARSF373-770 were co-transformed into *E. coli* ER-2566 for *in vivo* protein ligation [37]. The two precursor proteins were induced first with a final concentration of 0.09% arabinose for 0.5 h, followed by the addition of IPTG at a final concentration of 0.7 mM. The cells were induced for another 5.5 h before harvest. The His-tagged proteins were purified by IMAC using a 5-mL HisTrap HP column. The elution fractions were concentrated with an ultracentrifugal device with a molecular weight cut-off of 30 000.

Acknowledgements

We thank S. Ferkau and J. Mutanen for providing technical help with the protein and plasmid preparations. J.S.O. acknowledges the National Doctoral Programme in Informational and Structural Biology for financial support. A.S.A. was partly supported by the Viikki Doctoral Programme in Molecular Biosciences. We also thank the WeNMR project for the use of web portals, computing and storage facilities. This project was supported by the Academy of Finland (137995), the Sigrid Juselius Foundation and Biocenter Finland (NMR and mass-spectrometry facilities at the Institute of Biotechnology). The authors also thank Prof. P. Güntert for providing us the latest version of the program CYANA.

References

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

2 Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebl M & Stevens TH (1990) Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* **250**, 651–657.

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

4 Hall TM, Porter JA, Young KE, Koonin EV, Beachy PA & Leahy DJ (1997) Crystal structure of a hedgehog autoprocessing domain: homology between hedgehog and self-splicing proteins. *Cell* **91**, 85–97.

5 Perler FB (1998) Protein splicing of inteins and hedgehog autoproteolysis: structure, function, and evolution. *Cell* **92**, 1–4.

6 Perler FB (2002) InBase: the Intein Database. *Nucleic Acids Res* **30**, 383–384.

7 Amitai G, Belenkiy O, Dassa B, Shainskaya A & Pietrokovski S (2003) Distribution and function of new bacterial intein-like protein domains. *Mol Microbiol* **47**, 61–73.

8 Dassa B, Yanai I & Pietrokovski S (2004) New type of polyubiquitin-like genes with intein-like autoprocessing domains. *Trends Genet* **20**, 538–542.

9 Dassa B, Haviv H, Amitai G & Pietrokovski S (2004) Protein splicing and auto-cleavage of bacterial intein-like domains lacking a C'-flanking nucleophilic residue. *J Biol Chem* **279**, 32001–32007.

10 Dori-Bachash M, Dassa B, Peleg O, Pineiro SA, Jurkevitch E & Pietrokovski S (2009) Bacterial intein-like domains of predatory bacteria: a new domain type characterized in Bdellovibrio bacteriovorus. *Funct Integr Genomics* **9**, 153–166.

11 Volkmann G & Iwaï H (2010) Protein trans-splicing and its use in structural biology: opportunities and limitations. *Mol BioSyst* **6**, 2110–2121.

12 Lockless SW & Muir TW (2009) Traceless protein splicing utilizing evolved split inteins. *Proc Natl Acad Sci USA* **106**, 10999–11004.

13 Cheriyan M, Pedamallu CS, Tori K & Perler F (2013) Faster protein splicing with the *Nostoc punctiforme* DnaE intein using non-native extein residues. *J Biol Chem* **288**, 6202–6211.

14 Amitai G, Callahan BP, Stanger MJ, Belfort G & Belfort M (2009) Modulation of intein activity by its neighboring extein substrates. *Proc Natl Acad Sci USA* **106**, 11005–11010.
15 Minato Y, Ueda T, Machiyama A, Shimada I & Iwaï H (2012) Segmental isotopic labeling of a 140 kDa dimeric multi-domain protein CheA from *Escherichia coli* by expressed protein ligation and protein trans-splicing. *J Biomol NMR* **53**, 191–207.
16 Dawson PE, Muir TW, Clark-Lewis I & Kent SB (1994) Synthesis of proteins by native chemical ligation. *Science* **266**, 776–779.
17 Chang TK, Jackson DY, Burnier JP & Wells JA (1994) Subtiligase: a tool for semisynthesis of proteins. *Proc Natl Acad Sci USA* **91**, 12544–12548.
18 Mao H, Hart SA, Schink A & Pollok BA (2004) Sortase-mediated protein ligation: a new method for protein engineering. *J Am Chem Soc* **126**, 2670–2671.
19 Elleuche S & Pöggeler S (2010) Inteins, valuable genetic elements in molecular biology and biotechnology. *Appl Microbiol Biotechnol* **87**, 479–489.
20 Pietrokovski S (2001) Intein spread and extinction in evolution. *Trends Genet* **17**, 465–472.
21 Ding Y, Xu M, Ghosh I, Chen X, Ferrandon S, Lesage G & Rao Z (2003) Crystal structure of a mini-intein reveals a conserved catalytic module involved in side chain cyclization of asparagine during protein splicing. *J Biol Chem* **278**, 39133–39142.
22 Holm L & Rosenström P (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res* **38**, W545–W549.
23 Hiraga K, Soga I, Dansereau JT, Pereira B, Derbyshire V, Du Z, Wang C, Van Roey P, Belfort G & Belfort M (2009) Selection and structure of hyperactive inteins: peripheral changes relayed to the catalytic center. *J Mol Biol* **393**, 1106–1117.
24 Matsumura H, Takahashi H, Inoue T, Yamamoto T, Hashimoto H, Nishioka M, Fujiwara S, Takagi M, Imanaka T & Kai Y (2006) Crystal structure of intein homing endonuclease II encoded in DNA polymerase gene from hyperthermophilic archaeon *Thermococcus kodakaraensis* strain KOD1. *Proteins* **63**, 711–715.
25 Van Roey P, Pereira B, Li Z, Hiraga K, Belfort M & Derbyshire V (2007) Crystallographic and mutational studies of *Mycobacterium tuberculosis* recA mini-inteins suggest a pivotal role for a highly conserved aspartate residue. *J Mol Biol* **367**, 162–173.
26 Oeemig JS, Zhou D, Kajander T, Wlodawer A & Iwaï H (2012) NMR and crystal structures of the *Pyrococcus horikoshii* RadA intein guide a strategy for engineering a highly efficient and promiscuous intein. *J Mol Biol* **421**, 85–99.
27 Southworth MW, Yin J & Perler FB (2004) Rescue of protein splicing activity from a *Magnetospirillum magnetotacticum* intein-like element. *Biochem Soc Trans* **32**, 250–254.
28 Ellilä S, Jurvansuu JM & Iwaï H (2011) Evaluation and comparison of protein splicing by exogenous inteins with foreign exteins in Escherichia coli. *FEBS Lett* **585**, 3471–3477.
29 Oeemig JS, Aranko AS, Djupsjöbacka J, Heinämäki K & Iwaï H (2009) 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* **583**, 1451–1456.
30 Iwai H, Züger S, Jin J & Tam P-H (2006) Highly efficient protein trans-splicing by a naturally split DnaE intein from *Nostoc punctiforme*. *FEBS Lett* **580**, 1853–1858.
31 Aranko AS, Züger S, Buchinger E & Iwaï H (2009) In vivo and in vitro protein ligation by naturally occurring and engineered split DnaE inteins. *PLoS ONE* **4**, e5185.
32 Noren C, Wang J & Perler F (2000) Dissecting the chemistry of protein splicing and its applications. *Angew Chem Int Ed Engl* **39**, 450–466.
33 Shao Y, Xu MQ & Paulus H (1996) Protein splicing: evidence for an N-O acyl rearrangement as the initial step in the splicing process. *Biochemistry-Us* **35**, 3810–3815.
34 Cottingham IR, Millar A, Emslie E, Colman A, Schnieke AE & McKee C (2001) A method for the amidation of recombinant peptides expressed as intein fusion proteins in *Escherichia coli*. *Nat Biotechnol* **19**, 974–977.
35 Tori K & Perler FB (2011) Expanding the definition of class 3 inteins and their proposed phage origin. *J Bacteriol* **193**, 2035–2041.
36 Tori K, Cheriyan M, Pedamallu CS, Contreras MA & Perler FB (2012) The *Thermococcus kodakaraensis* Tko CDC21-1 intein activates its N-terminal splice junction in the absence of a conserved histidine by a compensatory mechanism. *Biochemistry-Us* **51**, 2496–2505.
37 Muona M, Aranko AS, Raulinaitis V & Iwaï H (2010) Segmental isotopic labeling of multi-domain and fusion proteins by protein trans-splicing in vivo and in vitro. *Nat Protoc* **5**, 574–587.
38 Muona M, Aranko AS & Iwaï H (2008) Segmental isotopic labelling of a multidomain protein by protein ligation by protein trans-splicing. *ChemBioChem* **9**, 2958–2961.
39 Farrow NA, Muhandiram R, Singer AU, Pascal SM, Kay CM, Gish G, Shoelson SE, Pawson T, Forman-Kay JD & Kay LE (1994) Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry* **33**, 5984–6003.
40 Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J & Bax A (1995) NMRPipe: a multidimensional spectral

processing system based on UNIX pipes. *J Biomol NMR* **6**, 277–293.
41 Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J & Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59**, 687–696.
42 Güntert P (2009) Automated structure determination from NMR spectra. *Eur Biophys J* **38**, 129–143.
43 Bertini I, Case DA, Ferella L, Giachetti A & Rosato A (2011) A Grid-enabled web portal for NMR structure refinement with AMBER. *Bioinformatics* **27**, 2384–2390.
44 Doreleijers JF, Vranken WF, Schulte CS, Markley JL, Ulrich EL, Vriend G & Vuister GW (2012) NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. *Nucleic Acids Res* **40**, D519–D524.
45 Nagai T, Ibata K, Park ES, Kubota M, Mikoshiba K & Miyawaki A (2002) A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nat Biotechnol* **20**, 87–90.
46 Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R & Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* **8**, 477–486.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site:

Fig. S1. [^{1}H,^{15}N]-HSQC spectrum of *Cth*BIL4 domain with sequence-specific assignments.

Fig. S2. ^{15}N relaxation analysis of *Cth*BIL4.

Fig. S3. MALDI-MS of the elution from IMAC.

Fig. S4. Time courses of *in vitro* ligation by split *Cth*-BIL4 at the C16 position.

Fig. S5. Western blotting analysis by anti-green fluorescent protein and anti-His antibodies and MALDI-MS of the elution from IMAC.