doi:10.1016/j.jmb.2012.04.029 J. Mol. Biol. (2012) 421, 85–99

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

NMR and Crystal Structures of the *Pyrococcus horikoshii* RadA Intein Guide a Strategy for Engineering a Highly Efficient and Promiscuous Intein

Jesper S. Oeemig [1], Dongwen Zhou [2], Tommi Kajander [1], Alexander Wlodawer [2] and Hideo Iwaï [1]*

[1]*Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki, P.O. Box 65, Helsinki FIN-00014, Finland*
[2]*Macromolecular Crystallography Laboratory, Frederick National Laboratory for Cancer Research, National Cancer Institute, Frederick, MD 21702, USA*

Received 19 March 2012; received in revised form 25 April 2012; accepted 27 April 2012
Available online 2 May 2012

Edited by K. Morikawa

Keywords:
protein engineering;
protein splicing;
extein;
trans-splicing;
segmental labeling

In protein splicing, an intervening protein sequence (intein) in the host protein excises itself out and ligates two split host protein sequences (exteins) to produce a mature host protein. Inteins require the involvement for the splicing of the first residue of the extein that follows the intein (which is Cys, Ser, or Thr). Other extein residues near the splicing junctions could modulate splicing efficiency even when they are not directly involved in catalysis. Mutual interdependence between this molecular parasite (intein) and its host protein (exteins) is not beneficial for intein spread but could be advantageous for intein survival during evolution. Elucidating extein–intein dependency has increasingly become important since inteins are recognized as useful biotechnological tools for protein ligation. We determined the structures of one of inteins with high splicing efficiency, the RadA intein from *Pyrococcus horikoshii* (*Pho*RadA). The solution NMR structure and the crystal structures elucidated the structural basis for its high efficiency and directed our efforts of engineering that led to rational design of a functional minimized RadA intein. The crystal structure of the minimized RadA intein also revealed the precise interactions between N-extein and the intein. We systematically analyzed the effects at the −1 position of N-extein and were able to significantly improve the splicing efficiency of a less robust splicing variant by eliminating the unfavorable extein–intein interactions observed in the structure. This work provides an example of how unveiling structure–function relationships of inteins offer a promising way of improving their properties as better tools for protein engineering.

**Corresponding author*. E-mail address: hideo.iwai@helsinki.fi.

Abbreviations used: COSY, correlated spectroscopy; ESRF, European Synchrotron Radiation Facility; HSQC, heteronuclear single quantum correlation; NOE, nuclear Overhauser effect; NOESY, NOE spectroscopy; *Npu*DnaE, DnaE intein from *Nostoc punctiforme*; *Pho*RadA, RadA intein from *Pyrococcus horikoshii*; PDB, Protein Data Bank; PTS, protein *trans*-splicing; *Sce*VMA, VMA intein from *Saccharomyces cerevisiae*; *Mxe*GyrA, GyrA intein from *Mycobacterium xenopi*; *Mja*KlbA, KlbA intein from *Methanococcus jannaschi*; *Ssp*DnaE, DnaE intein from *Synechocystis sp.* strain PCC6803; *Ssp*DnaB, DnaB intein from *Synechocystis sp.* strain PCC6803; TOCSY, total correlation spectroscopy.

Introduction

Protein splicing is a unique self-catalytic reaction in which an internal protein sequence (intein) removes itself from the host protein, concomitantly ligating the two flanking sequences (exteins) with a peptide bond.[1] Inteins have been identified in bacteria, archaea, and unicellular eukaryotes, with more than 550 putative inteins currently registered in the intein database (InBase).[2] Protein splicing can also occur in *trans* via split inteins, which could be naturally occurring or artificially engineered.[3] Protein *trans*-splicing (PTS) has become an important tool for protein ligation *in vivo* and *in vitro*, enabling useful biotechnological applications that include segmental isotopic labeling, site-specific modifications, and protein cyclization.[4–7] However, extensive applications of protein ligation by PTS have been often hindered by poor tolerance of splicing junction sequences and their ligation efficiency.[8,9]

A well-understood typical mechanism of protein splicing involves four concerted steps: (1) N–S(O) acyl shift, (2) *trans*-(thio)esterification, (3) Asn cyclization, and (4) S(O)–N acyl shift.[10] Additionally, noncanonical inteins that do not follow this standard mechanism have also been reported.[11,12] The first residue of inteins is usually a Ser or Cys that catalyzes N–S(O) migration at the N-terminal junction to initiate protein splicing. Class II and class III inteins contain as the first residue Ala that cannot initiate the first step of N–S(O) shift. These inteins utilize different pathways to achieve protein splicing. Structural studies have suggested that specific roles for other residues in the proximity of the catalytic residues differ significantly between different inteins.[13–16] Variations in the catalytic mechanism seem to be better tolerated in protein splicing than for other enzymes, possibly because intein is a single-turnover enzyme with fused substrates. On the other hand, neighboring extein residues near the splicing junctions that are not directly involved in the catalytic reaction could significantly influence the splicing efficiency of inteins.[8,17,18] This extein dependency could restrict practical applications of inteins because it makes it hard to predict whether protein splicing reaction would result in productive splicing or off-pathway cleavages in foreign contexts.[8,19,20] Therefore, it is of considerable significance for the advancement of biotechnological applications of inteins to understand the atomic details of extein dependency. Previously, we identified RadA intein from *Pyrococcus horikoshii* (*Pho*RadA) as an intein with highly efficient splicing activity, similar in its efficiency to the *cis*-splicing variant of naturally split DnaE intein from *Nostoc punctiforme* (*Npu*DnaE).[21] The high splicing efficiency and its relatively small size make *Pho*RadA intein attractive for developing novel biotechnological tools. In order to elucidate the structural basis of the activity of *Pho*RadA intein, we have determined its structures by NMR and X-ray crystallography. These structural results have led us to further redesign of this intein.

Results and Discussion

Inteins with robust splicing activity are indispensable for applications of PTS such as protein ligation. *Npu*DnaE intein was found to be one of the highly efficient split inteins widely used for PTS.[8,21] For the applications of multi-fragment ligation with PTS, it is critical to have at least one more robust split intein that is orthogonal to naturally split *Npu*DnaE intein, although a few approaches have been reported to circumvent the cross-activity between split inteins.[22,23] We determined the three-dimensional structure of *Pho*RadA intein in order to gain better understanding of its mode of activity and to provide a basis for creating new biotechnological tools.

The NMR solution structure of *Pho*RadA intein

As a first step to elucidate the NMR structure of the full-length *Pho*RadA intein, ^{13}C, ^{15}N-labeled *Pho*RadA intein was recombinantly produced in *Escherichia coli*. This construct contains the intein with a mutation of the first residue (C1A), a two-residue C-extein with a mutation of the first residue of C-extein (T+1A), and no N-extein (Fig. 2). The resonance assignments of *Pho*RadA intein with two C-extein residues were nearly completed by conventional triple-resonance experiments, with 94.8% of all the atoms except for the C-extein and a few residues (Fig. 1). Based on the assignments, the nuclear Overhauser effect (NOE) peaks were iteratively analyzed by the automated algorithm implemented in CYANA 3.0 in order to determine the structure.[24] The final solution structure of *Pho*RadA intein contains predominantly β-sheets with a short helix between residues 21 and 31 and a short 3_{10} helix (Fig. 2). The solution NMR structures were well defined and highly consistent, with a backbone root-mean-square deviation (RMSD) of 0.61 Å for residues 1–172, excluding the two C-extein residues (Table 1). The overall fold of *Pho*RadA intein belongs to the HINT (*h*edgehog/*int*ein) domain superfamily (Fig. 3).[27] As expected from its primary structure, *Pho*RadA intein is a mini-intein lacking an endonuclease domain, the biological function of which, other than catalyzing protein splicing, is still largely unknown. The most highly disordered regions of the structure are at residues 121–133 and 173–174, where their resonance assignments are missing and only a few NOEs could be identified (Fig. 3). The heteronuclear NOE and ^{15}N relaxation data also indicate that the loop around residues 121–133 is disordered (Supplemental Fig. 1).

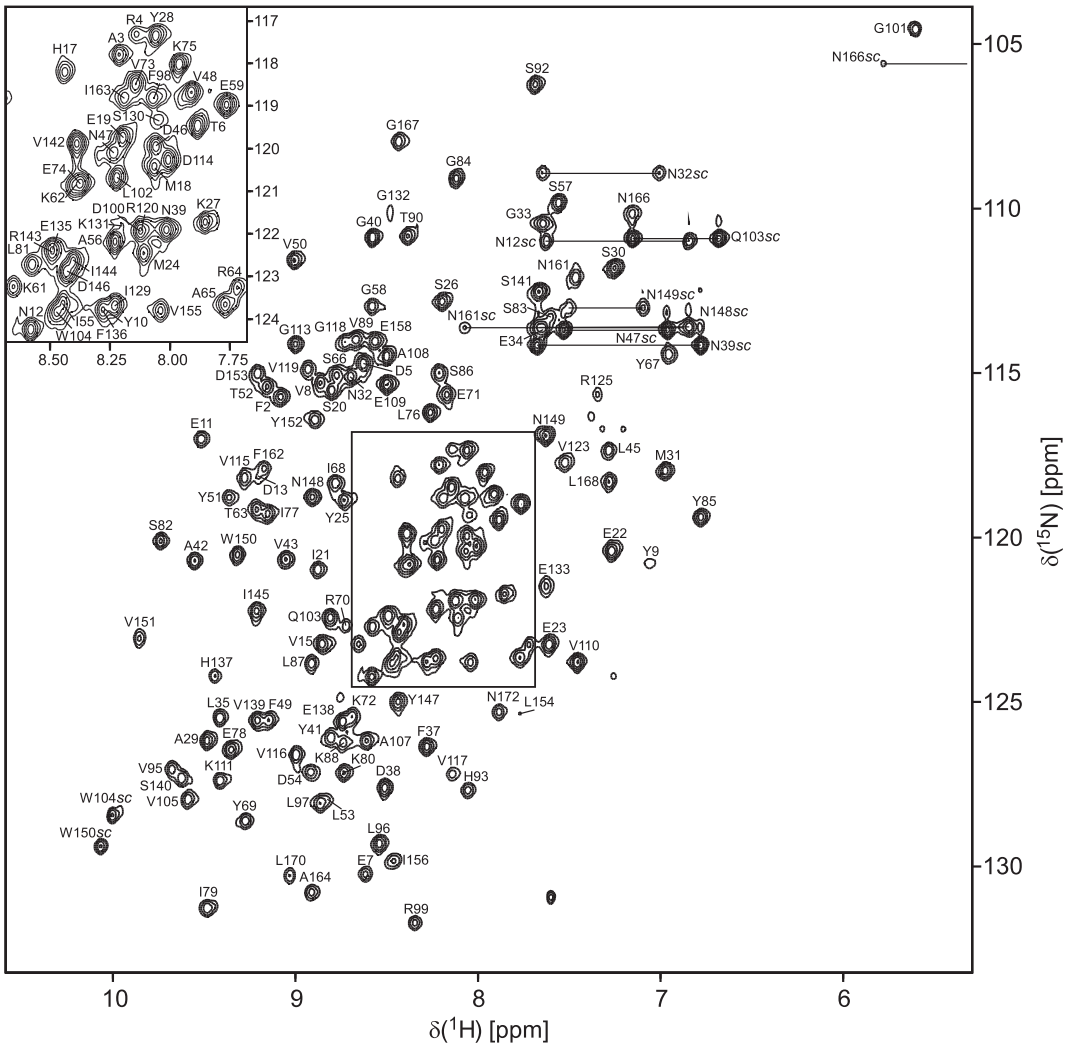

Fig. 1. [^{1}H,^{15}N]-HSQC spectrum of 0.4 mM *Pho*RadA intein (pH 6.0) recorded at ^{1}H frequency of 800 MHz. Sequence-specific assignments are labeled. The side-chain assignments are indicated by "sc". Side-chain resonances of Asn and Gln are connected by horizontal lines.

The crystal structures of *Pho*RadA intein

The construct of *Pho*RadA intein used for the NMR structure determination contains the C1A mutation in the intein and a two-residue C-extein with a mutation of Thr to Ala (T+1A), aimed at prevention of splicing and cleavage (Fig. 2). However, we could not define the conformation around the C-extein due to the lack of sufficient assignments for these residues. In order to find out if more information about this region could be gleaned from a crystal structure, we crystallized the construct of *Pho*RadA intein identical with the one used in the NMR study.[28] Its crystal structure was determined at 1.75 Å resolution by the structure and density optimized molecular replacement implemented in Rosetta,[29] using the NMR structures as starting models (Fig. 4a). The final model of *Pho*RadA intein contains 2 polypeptide chains, 1 Hepes buffer molecule, 4 glycerol molecules, and 124 water molecules in the asymmetric unit. In the polypeptide chain A, 172 amino acid residues have been modeled, although loop regions 13–14, 55–57, and 125–129 show only very weak electron density. The last two residues of the C-extein region could not be modeled due to lack of electron density. In the polypeptide chain B, 169 residues have been modeled, including the loop region 12–15, which shows weak electron density. However, the loop region 127–129 was not included in the model of this chain due to complete lack of electron density. In addition, density was missing or incomplete for a number of side chains. The quality of the Ramachandran plot of the final model of *Pho*RadA intein was evaluated by PROCHECK,[30] which shows 93.3%, 5.4%, and 1.3% of all residues falling into

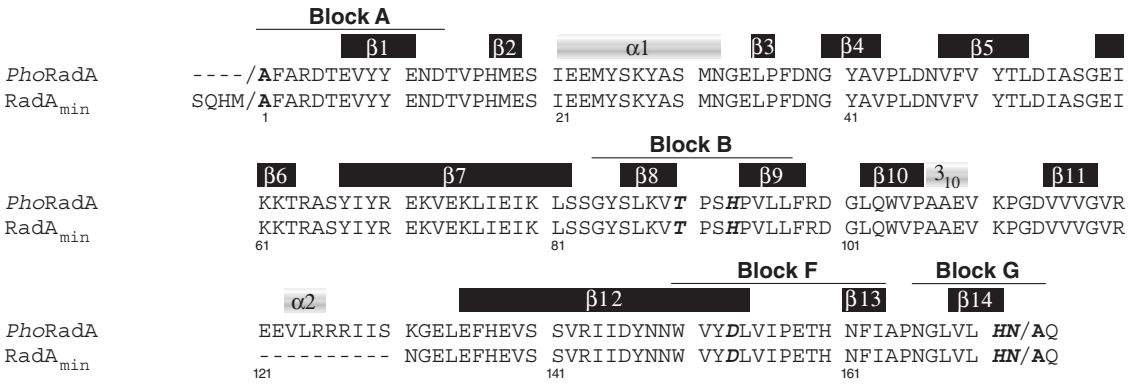

Fig. 2. Primary structures of *Pho*RadA intein and minimized *Pho*RadA intein ($RadA_{min}$) used for the structure determination. Secondary structures based on the determined NMR model are indicated above the protein sequence. Conserved intein sequence regions of blocks A, B, G, and H are marked. The mutated active-site residues are in boldface. Highly conserved His and Thr in block B, Asp in block F, and His and Asn in block G in the active site are in boldface and italic (see the text).

the most favored, additionally allowed, and generously allowed regions, respectively. Of the three residues in the latter category, two are located in the loops with poor electron density, but Arg99 of molecule A is located in very well defined density. Both the main chain and the side chain of this residue make a number of hydrogen bonds, which are most likely responsible for its non-ideal torsion angles. The conformations of the two molecules in the asymmetric unit of the crystals of *Pho*RadA intein are very similar, with an RMSD of 0.254 Å for the 138 superimposed C^{α} atoms (Fig. 4a). The most significant deviation is present in the loop region containing residues 11–17 (Fig. 4a). In chain B, the loop is longer than its counterpart in chain A, since it includes residues 11, 16, and 17, which are part of β-strands in chain A. In the loop region 125–129, chain B is more disordered, and three residues (Arg127, Ile128, and Ile129) are not included in the final model due to complete lack of electron density.

Table 1. Statistics of the NMR structure calculation of *Pho*RadA intein

Quantity	Value
NOE upper distance limits	3515
Short-range NOE ($I-j\leq 1$)	1654
Medium-range NOE ($1<i-j<5$)	425
Long-range NOE ($i-j\geq 5$)	1436
Dihedral angels[a]	278
Residual CYANA target function (Å^2)	1.14 ± 0.08
Residual NOE violation	
Number ≥ 0.3 Å	3 ± 2
Maximum (Å)	0.67 ± 0.24
Residual dihedral angle violation	
Number ≥ 5°	6 ± 2
Maximum (°)	12.28 ± 2.96
AMBER energies (kcal/mol)	
Total	−6518 ± 24.18
van der Waals	−1402 ± 15.8
Electrostatic	−11,520 ± 330
RSMD from ideal geometry	
Bond length (Å)	0.0229 ± 0.0001
Bond angels (°)	2.120 ± 0.015
RMSD to mean coordinate	
Backbone 1–172 (Å)	0.61 ± 0.10
Heavy atoms 1–172 (Å)	1.09 ± 0.16
Ramachandran plot statistics[b] (%)	
Most favored regions	91.3
Additional allowed regions	8.2
Generously allowed regions	0.5
Disallowed regions	0.0

[a] TALOS+-derived angle predictions.[25]
[b] Derived by PROCHECK-NMR.[26]

Comparison between the NMR and crystal structures of *Pho*RadA intein

The NMR structure of *Pho*RadA intein contains all the 174 amino acid residues of the construct, including the last two residues at the C-terminus that are part of the extein, which, however, were not well defined due to the lack of NOE constraints. This C-extein region was also not seen in the crystal structures of *Pho*RadA intein, confirming conformational flexibility of this region in both structures. The RMSD between the NMR structure and chain A of the crystal structure is 0.905 Å for 142 superimposed residues, whereas the RMSD between the NMR structure and chain B is 0.808 Å for 135 superimposed residues. The secondary structure elements are generally the same in both crystal and NMR structures. Relatively large deviations mainly occur at the loop regions, and the most significant one is located in the loop 120–133 (Fig. 4b). The displacement between main chains in this region is around 6.8 Å. Not surprisingly, this loop is the most flexible region of the whole structure. Nevertheless, the overall deviation between the crystal and NMR-derived coordinates is on par with other protein structures investigated by these two techniques.[31]

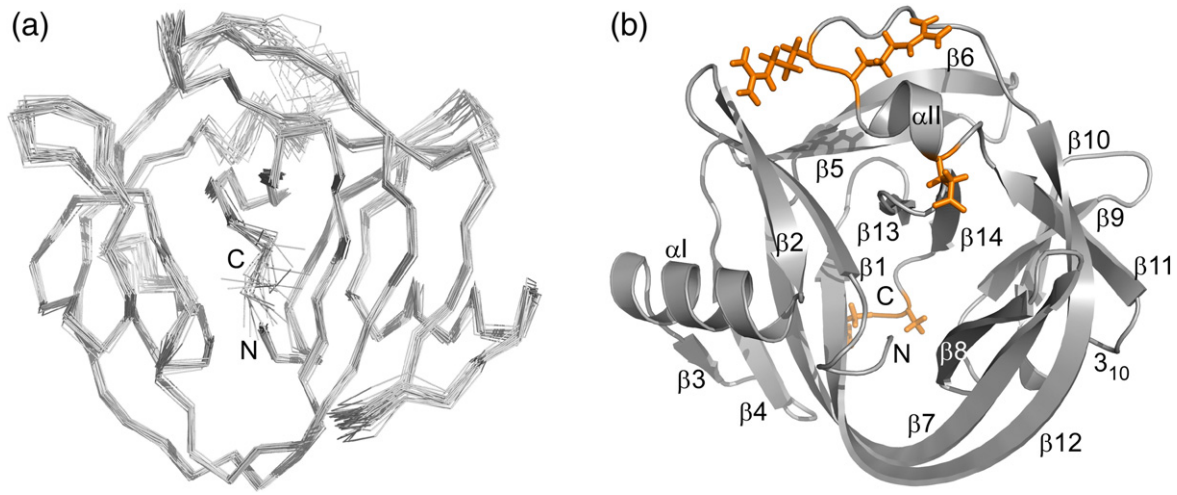

Fig. 3. The NMR solution structure of *Pho*RadA intein. (a) The 20 superimposed NMR structures. (b) Ribbon presentation of the NMR structure labeled with the secondary structures. Residues that were not assigned are shown in stick model and colored orange. The N-terminus and the C-terminus are indicated.

Minimizing *Pho*RadA intein

Both NMR and crystal structures revealed a disordered loop that included residues 120–133. The loop is located between blocks B and F, where many inteins contain an endonuclease domain that is presumably important for horizontal gene transfer.[1,3] The presence of disorder prompted us to test if this loop could be shortened without loss of function. We removed 10 residues between β11 and β12 and replaced Lys131 by Asn to better accommodate a turn (Fig. 2 and 5a). The newly minimized *Pho*RadA intein (RadA$_{min}$ intein) was tested with a model system using two B1 domains of immunoglobulin binding protein G (GB1) as exteins for *cis*-splicing (Fig. 5c).[21] Expression of the *cis*-splicing precursor protein bearing the engineered RadA$_{min}$ intein predominantly produced the *cis*-spliced

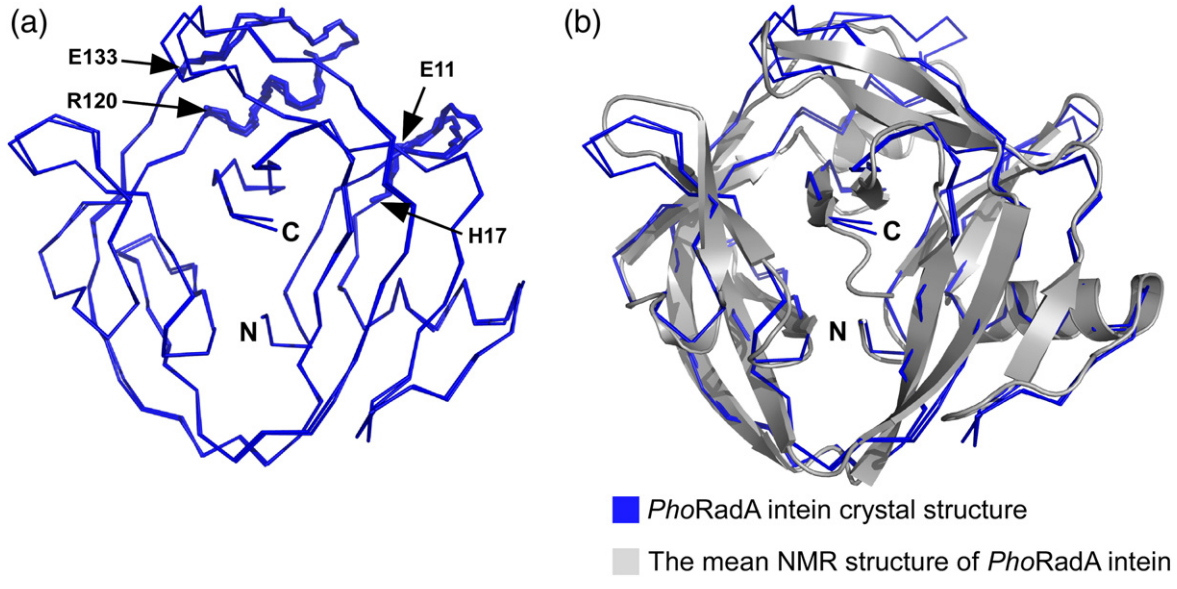

Fig. 4. The crystal structure of *Pho*RadA intein. (a) Superposition of the two *Pho*RadA intein models in the asymmetric unit. The loop regions containing residues 11–17 and 120–133 are shown in stick representation with numbering for the beginnings and the ends. (b) The mean NMR model in ribbon representation (gray) is superimposed with the two models of the *Pho*RadA intein crystal structures (blue). The N-terminus and the C-terminus are indicated. Backbone atoms of residues 1–126 and 130–172 are used for the superposition of the structures.

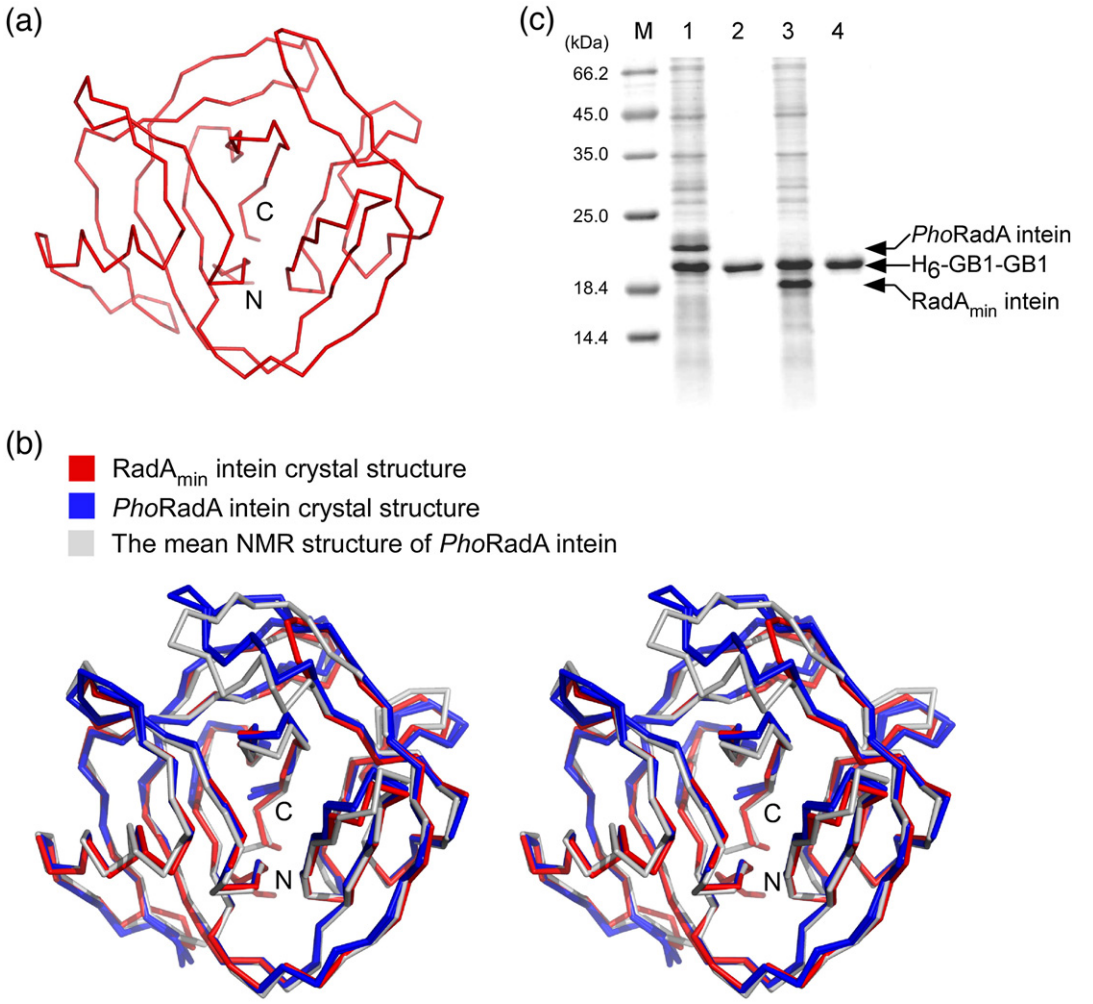

Fig. 5. Minimized *Pho*RadA intein ($RadA_{min}$) structure. (a) Structure of $RadA_{min}$ intein (red). (b) A stereo view of the superposition of $RadA_{min}$ intein structure (red), crystal structures of *Pho*RadA intein (blue), and the mean NMR structure of *Pho*RadA intein (gray). The N-terminus and the C-terminus are indicated. (c) SDS-PAGE analysis of *cis*-splicing of *Pho*RadA intein and $RadA_{min}$ intein. M indicates protein marker. Lane 1, the total cell lysate of the cells expressing *cis*-splicing precursor with *Pho*RadA intein; lane 2, elution fraction from the Ni-NTA column loaded with the cell lysate expressing the *Pho*RadA intein precursor; lane 3, the total cell lysate of the cells expressing the precursor bearing $RadA_{min}$; lane 4, elution fraction from the Ni-NTA column loaded with the cell lysate expressing the precursor containing $RadA_{min}$ intein.

product (H_6-GB1-GB1), confirming that $RadA_{min}$ intein is highly functional (Fig. 5c). Both precursors bearing *Pho*RadA and $RadA_{min}$ inteins produced only the spliced product of H_6-GB1-GB1 and the excised intein after 4 h of the induction (Fig. 5c, lanes 1 and 3). The elution fractions from the Ni-NTA column contained only the *cis*-spliced product (H_6-GB1-GB1), but neither the precursor nor cleaved products were present in the elution fraction (Fig. 5c, lanes 2 and 4). Thus, *cis*-splicing of the engineered $RadA_{min}$ intein is very efficient and indistinguishable from that of *Pho*RadA intein, suggesting that the removed disordered loop is a mere remnant from the endonuclease domain that was lost during evolution.

The crystal structure of $RadA_{min}$ intein

The structure of the newly designed $RadA_{min}$ intein was verified by X-ray crystallography. We crystallized the protein containing both four N-extein residues and two C-extein residues, with the alanine mutations at the first residues of the intein and C-extein (Fig. 2), and we obtained diffraction data extending to 1.58 Å resolution. The better diffraction of $RadA_{min}$ intein compared to the parent enzyme could be due to the removal of the disordered loop. The final model of $RadA_{min}$ intein contains a single polypeptide chain and 251 water molecules in the asymmetric unit (Fig. 5a). All 168 amino acid residues present in the expression

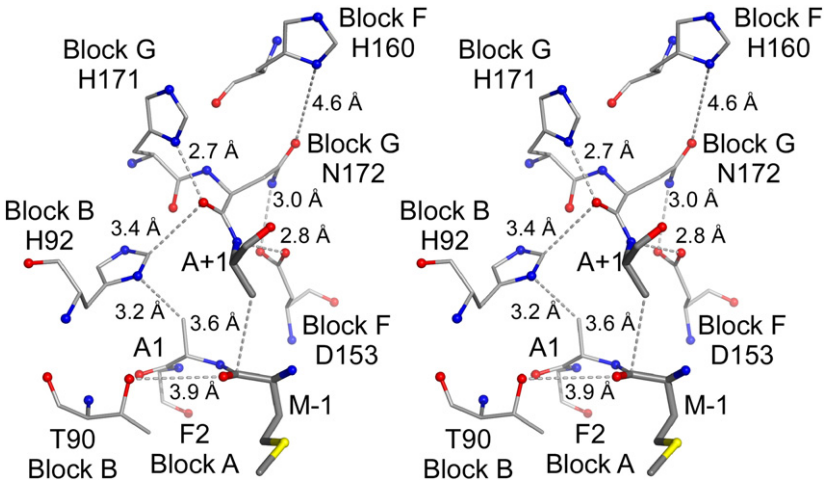

Fig. 6. The structure of the splicing site of $RadA_{min}$ intein. A stereo view of residues located in the vicinity of the splicing site. Distances between atoms between the conserved residues are indicated with broken lines. N, O, and S atoms are indicated with blue, red, and yellow spheres, respectively. The residues from the N- and C-exteins are colored dark gray. Residue numbers are labeled with the locations in the conserved intein motif (blocks A, B, F, and G). Only backbone atoms of Phe2 are shown.

construct have been modeled, and only residues 13, 82, and 83, which are all located in the loop regions, show relatively weak electron density. The electron density for the four extein residues at the N-terminus and for two additional residues at the C-terminus that are not part of the intein is of excellent quality, and the conformation of the termini is not in doubt (Fig. 7). The Ramachandran plot of the final $RadA_{min}$ model shows 92.4%, 6.2%, 0.7%, and 0.7% of all residues falling into the most favored, additionally allowed, generously allowed, and disallowed regions, respectively. Not surprisingly, the only residue that falls into the disallowed region, Asp13, is located in a flexible loop with very weak electron density.

Comparison between *Pho*RadA and $RadA_{min}$ inteins

Whereas the crystal structure of *Pho*RadA intein contains two molecules in the asymmetric unit, there is only one molecule in the asymmetric unit of $RadA_{min}$ structure. The RMSD between $RadA_{min}$ and *Pho*RadA intein chain A is 0.493 Å for the 130 superimposed residues, and the RMSD between $RadA_{min}$ and *Pho*RadA intein chain B is 0.541 Å for the 133 superimposed residues (Fig. 5c). The loop region 120–133 of the native *Pho*RadA intein structure, which includes a short α-helix, is quite flexible as demonstrated by the weak electron density and comparatively high temperature factors. In the $RadA_{min}$ intein structure, because of the 10-residue deletion, the remaining four residues became quite stable, and all the residues have complete electron density, suggesting that the designed deletion did not disturb the rest of the structure (Fig. 5b). As this loop region is poorly ordered and dispensable for the splicing, it is an excellent candidate for introducing a split site for developing new split inteins.[32]

Structure of the catalytic site

The crystal structure of $RadA_{min}$ intein allows us to investigate at atomic resolution the catalytic site together with the N- and C-extein residues. Unlike the *cis*-conformation of the N-scissile peptide bond reported in GyrA intein from *Mycobacterium xenopi* (*Mxe*GyrA),[13] the scissile peptide bond between Met –1 and Ala1 of $RadA_{min}$ intein is in the usual *trans*-conformation. In the structure of the mutant of VMA intein precursor from *Saccharomyces cerevisiae* intein precursor (X10SSS), the carbonyl oxygen and the first residue O^{γ} are 3.1 Å apart.[33] This distance is short enough for a nucleophilic attack between them, suggesting the formation of a five-membered ring thiazolidine intermediate during the N–S acryl shift.[33] The existence of the suggested thiazolidine intermediate was recently confirmed with DnaB intein from *Synechocystis* sp. strain PCC6803 (*Ssp*DnaB).[34] In the RadA intein structures, the first residue (Cys1) is replaced by Ala; therefore, it is not possible to measure the same distance as in the mutant of *Sce*VMA intein. However, modeling of Cys1 based on the experimental structure of $RadA_{min}$ intein indicates that the corresponding distance could be as short as 3.0 Å, similar to the mutant of *Sce*VMA intein, indicating that *Pho*RadA intein may utilize the same tetrahedral intermediate during the N–S acyl shift (Fig. 6). The structure of $RadA_{min}$ intein also indicates a very short distance between the carbonyl carbon of the N-scissile peptide and the side chain of the +1 residue that is responsible for the second step of *trans*-esterification (Table 2 and Fig. 6). The close proximity of these key atoms suggests that *trans*-esterification could take place without large conformational changes after the first step N–S acyl rearrangement. This structural feature might account for the highly efficient *cis*-splicing of *Pho*RadA intein.

An aspartate residue in block F of *Pho*RadA intein, located in the vicinity of the catalytic site, is highly conserved among inteins. That residue was found to

Table 2. Intein structures with N- and C-extein residues

Intein	PDB	Resolution (Å)	Wild-type N-extein	Wild-type C-extein	Model N-extein	Model C-extein	Distance C′(−1)–S(O)$^{\gamma}$(+1) (C′–C$^{\beta}$) (Å)	Mutations and additional remarks
*Ssp*DnaE[14]	1ZDE	1.95	FAEY	CFNK	FEKY/A	A/CFN	8.0 (7.8)	C1A, N158A, C+1 is coordinated by Zn ion
*Ssp*DnaB[13]	1MI8	2.0	RESG	SIWQ	SG/A	A/SI	8.0 (7.8)	C1A, N429A, endonuclease domain is removed
*Sce*VMA[33]	1JVA	2.1	IYVG	CGER	VG/S	S/SGER	3.8 (4.1)	C1S, H79N, N454S, C+1S
*Sce*VMA[35]	1EF0	2.1	IYVG	CGER	KLEG/A	A/CGER	9.1 (7.8)	C1A, N454A, C+1 is coordinated by Zn ion
*Mja*KlbA[16]	2JMZ	NMR	GHDG	CSGT	GHDG/A	A/SSGT	8.1 (9.3)	N168A, C+1S
RadA$_{min}$	4E2U	1.58	GSGK	TQLA	SQHM/A	N/AQ	— (3.7)	C1A, T+1A, residues 121–130 are removed

play a pivotal role during the splicing in several inteins (Fig. 6).[14,16,36] The side chain of Asp153 in block F forms hydrogen bonds with the backbone of the +1 residue, as well as with the side chain of the last residue of the intein, (Asn172), which catalyzes the cleavage of the branched intermediate after *trans*-esterification step. The interaction between Asn172 and Asp153 keeps the side chain of Asn172 away from the C-scissile bond, thereby preventing the third step of Asn cyclization. This structure implies that conformational rearrangement to induce Asn cyclization after the *trans*-esterification step is necessary. This particular interaction has not been observed with other intein structures that contained extein sequences, since the last Asn residues were replaced by Ala or Ser in these structures (Table 2). The orientation of Asn172 observed in the structure could be important for the understanding of the required structural rearrangements necessary at the individual steps during the protein splicing.

Extein–intein interactions

Since our aim is to create more promiscuous inteins with respect to the splicing junctions, we are particularly interested in understanding the roles of extein sequences in protein splicing. In the structures of full-length *Pho*RadA intein determined by X-ray crystallography and NMR spectroscopy, it was not possible to define the precise conformation of the C-extein region. The results obtained by both techniques indicate the existence of conformational flexibility in this region of this protein. We have introduced a mutation from Thr to Ala at the +1 position of C-extein and also removed the N-extein sequence, which might have increased the flexibility of the extein region. A flexible C-extein structure was also reported in another intein structure.[16] Although local mobility is suggested to be important for several steps of chemical reactions to take place, it is still not clear whether such flexibility is functionally essential or simply a result of shortening of the C-extein or removal of the N-extein. On the other hand, we could precisely define the conformation of the extein sequences in the case of RadA$_{min}$ intein. Despite the increasing number of intein structures deposited in the Protein Data Bank (PDB), only a few of them have been determined with both N- and C-extein residues modeled (Table 2). Among these intein structures, the distances between the carbonyl carbon atom of the N-terminal scissile bond and S$^{\gamma}$ (or O$^{\gamma}$) atom of the first residue of C-extein are often quite long (8–9 Å), except for the mutant (X10SSS) of *Sce*VMA intein (PDB code: 1JVA) where it is ~4 Å.[33] The RadA$_{min}$ intein structure is similar to that of the latter intein as the −1 and +1 residues are closer in the three-dimensional structure ("closed" conformation). The ψ angle of the first residue of RadA$_{min}$ intein is −35°, which is significantly different from ψ angles of 159° for *Ssp*DnaE (1ZDE), 151° for *Ssp*DnaB (1MI8), and 127 ± 7° for *Sce*VMA (1JVA).[14,15,33] Two of the previously reported inteins with the open conformation contain a zinc ion near the active site. It has been demonstrated that zinc can inhibit protein splicing of inteins, suggesting that its coordination observed in the crystal structures could be responsible for the longer distance between the −1 and +1 residues, locking them in the open conformation.[15] The crystal structures using several variants of inteins probably represent snapshots of the various conformations that may occur during protein splicing, which certainly requires structural rearrangements to accommodate a few different chemical reaction steps at the same site.

Implication for N-extein dependency

Based on the N-extein confirmation seen in the structure of RadA$_{min}$ intein, it is now possible to scrutinize the structural details of the extein–intein interactions of the *Pho*RadA intein. To elucidate the structural basis of N-extein dependency of *Pho*RadA intein, we compared *cis*-splicing of *Pho*RadA intein with 20 different amino acid types at the −1 position of N-extein using a model system (Fig. 8a). *Pho*RadA intein not only has highly efficient splicing activity but also accepts many side chains at the −1 position

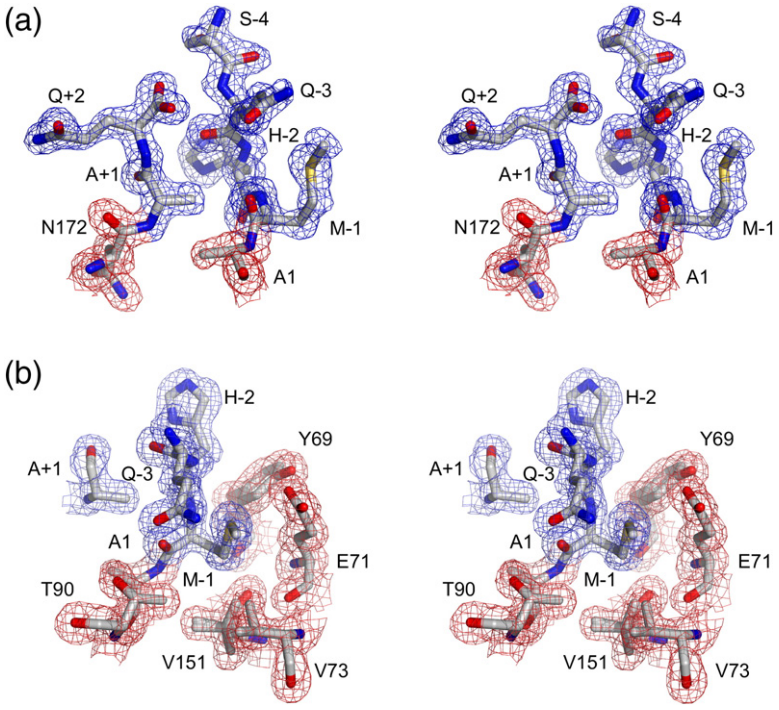

Fig. 7. Electron density map of $RadA_{min}$ intein and the extein residues. (a) A stereo view of the map of the extein residues (blue mesh) and the first and last residues of $RadA_{min}$ intein (red mesh). (b) A stereo view of the map showing the residues within 4 Å of the last residue of N-extein (M-1). The electron density map is contoured at 1 σ.

without significant loss of function (Fig. 8b and Supplemental Fig. 2), making this intein very attractive for biotechnological applications. This comparison revealed that proline, amino acids with negative charges (Glu and Asp), β-branched amino acids (Val, Ile, and Thr), and amino acids with smaller side chains (Gly and Ala) have lower *cis*-splicing efficiency. It is not surprising that proline at the −1 position has the lowest splicing activity because the scissile peptide is chemically different from other peptide bonds as it lacks H^N hydrogen atom. In the crystal structure of $RadA_{min}$ intein, the side chain of the −1 residue makes contacts with Glu71, Val73, and Val151 (Fig. 7). The native N-extein of *Pho*RadA intein has Lys at the −1 position, and its positively charged side chain is making a charge–charge interaction with the negatively charged Glu71. Whereas all other residues with longer side chains (van der Waals volume of >70 $Å^3$) retained good splicing activity, residues with smaller van der Waals volumes, such as Ala (67 $Å^3$) and Gly (48 $Å^3$), slightly lowered the splicing efficiency.[37] This suggests that van der Waals contacts are important for the proper alignment of the active-site atoms to direct the reaction toward the productive splicing pathway. It can be postulated that β-branched amino acids (Val, Ile, and Thr) disturb the precise arrangement of the active-site groups due to their bulky side chains, lowering the productive *cis*-splicing. It is noteworthy that β-branched amino acids are also similarly unfavorable for *trans*-esterification in native chemical ligation.[38] It might be of interest to investigate whether the effect of β-branched residues for both reactions is due to a similar mechanism.

Engineering a promiscuous intein at the N-junction

We hypothesized that the negative charge of Glu71 could be the cause for lowering splicing efficiency of the two variants bearing Glu or Asp at the −1 position (the E-1 and D-1 variants), presumably due to unfavorable negative charge interactions. To test our hypothesis, we introduced a mutation of E71T in *Pho*RadA intein to remove the negative charge. The mutant E71T indeed significantly improved *cis*-splicing of the E-1 variant from about 40% to >90% by lowering the amount of unspliced precursor, as well as by suppressing the cleavage at the C-terminal junction (Fig. 8c, lanes 2 and 4). Moreover, the E71T mutant of *Pho*RadA intein also retains high splicing activity with the native −1 residue of Lys (Fig. 8c, lane 5). This observation confirms that the negative charge of Glu71 is responsible for the lower efficiency of the E-1 variant. In the case of the D-1 variant, the *cis*-splicing was not improved at all, despite the same negative charge of the side chain (Fig. 8c, lanes 1 and 3). Unlike the E-1 variant, the D-1 variant resulted in mostly N-cleavage, as can be seen in the elution fraction containing significant amounts of N-cleaved product of H_6-GB1 (Fig. 8c, lanes 1 and 3). This result suggests that the lower efficiency of the D-1 variant is not due to the same mechanism as the E-1 variant. A similar N-cleavage with Asp

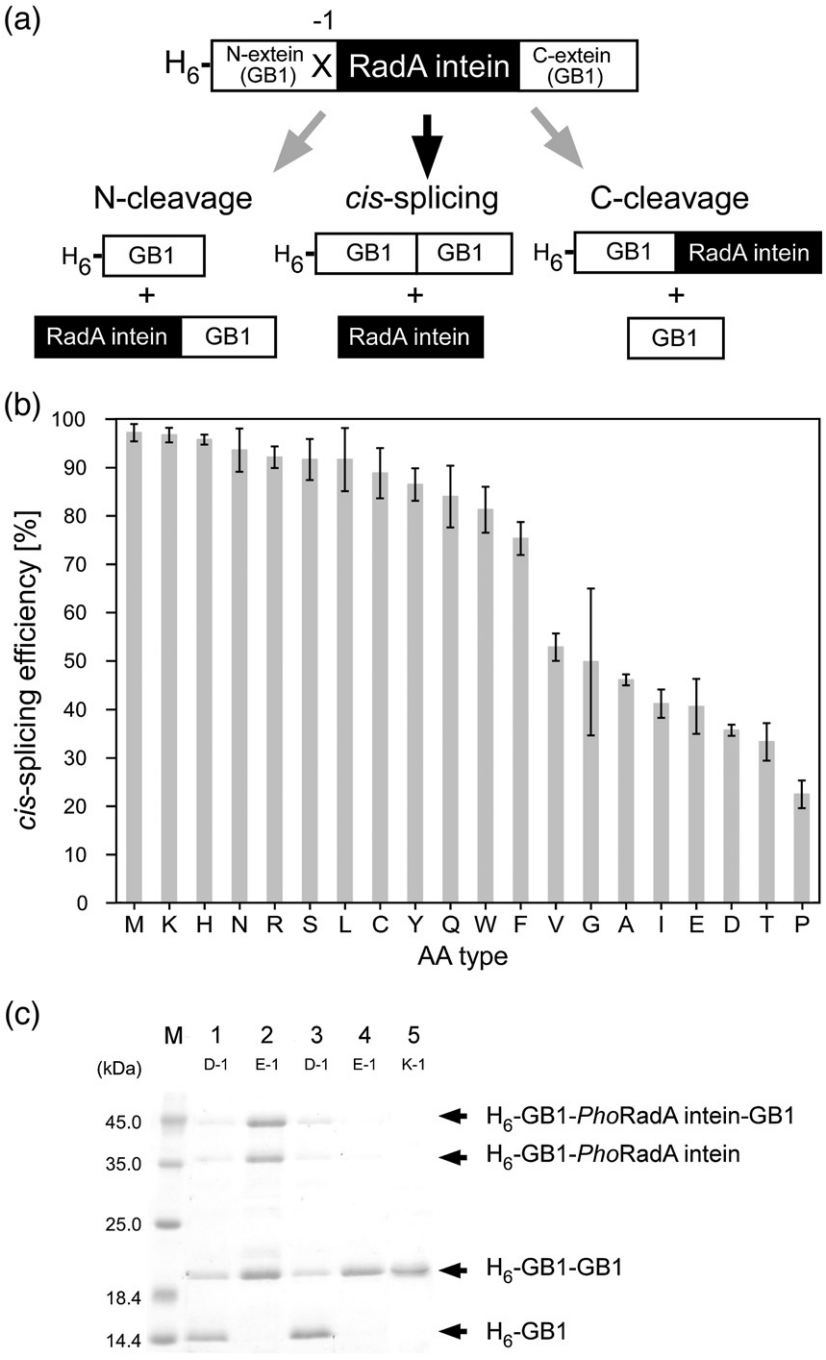

Fig. 8. N-extein dependency of *Pho*RadA intein. (a) A model system used for evaluating *cis*-splicing efficiency. X stands for 20 amino acids at the −1 position tested for the comparison. The precursor protein could result in either on-pathway *cis*-splicing, off-pathway N-cleavage, or off-pathway C-cleavage. (b) A plot of *cis*-splicing efficiency (%) against 20 amino acid types at the −1 position. The error bars were obtained from the analysis of three independent experiments. (c) Comparison of protein splicing efficiency of *Pho*RadA intein and the E71T variant. The purified fractions from the cells expressing *cis*-splicing precursors by Ni-NTA column were analyzed by SDS-PAGE. M indicates protein markers. Lane 1, D-1 variant of *Pho*RadA intein; lane 2, E-1 variant; lane 3, D-1 variant with E71T mutation in the *Pho*RadA intein; lane 4, E-1 variant with the E71T mutation; lane 5, K-1 variant with the E71T mutation.

at the −1 position was reported for *Mxe*GyrA, *Ssp*DnaE, and *Sce*VMA inteins.[17,39,40] It is known that an aspartic acid in a polypeptide chain could hydrolyze at its C-terminal side under acidic conditions due to a nucleophilic attack of the side-chain carboxyl group, depending on peptide sequences.[41] The N-cleavage of the D-1 variant is possibly due to a similar involvement of the aspartate side chain at the −1 position that cleaves the N-scissile peptide bond or the thioester after N–S acyl shift. The InBase lists dozens of inteins with Asp at the −1 position,[2] and it is likely that they utilize another structural mechanism to prevent the N-cleavage.

In summary, we have determined the solution and crystal structures of *Pho*RadA intein. One of the structures allowed us to investigate the intein–extein interactions for understanding the structural basis of the N-extein dependency. Inteins are often considered to be mere parasitic proteins inserted into host

proteins, providing no evident benefit to the host organisms. From the perspective of parasites, strong mutual extein–intein interactions could be less beneficial for intein spreads as it might not function at different insertion points.[3] The structures of *Pho*RadA inteins revealed the defined interactions between the −1 position of N-extein and the intein, as well as conformational flexibilities in the extein residues near the splicing junction. This interaction between N-extein and *Pho*RadA intein can play an important role for the highly efficient protein splicing, as observed with the 20 variants at the −1 position. Starting with the structural results, we demonstrated that a mutation in the intein could significantly improve protein splicing efficiency of the less robust E-1 variant, suggesting that *Pho*RadA intein has acquired mutations after the insertion into the host protein but retained the efficient splicing activity in the context of the insertion point of the host protein. Ancestral inteins that spread to different host proteins might have been less dependent on the splicing junction sequences except for the first residue of C-extein that needs to be Cys, Ser, or Thr.[3] In the intein database, all 20 common residue types except for Trp have been reported at the −1 position of the putative inteins. Although not all inteins have been tested for protein splicing, it is likely that inteins could tolerate all naturally occurring amino acids at least at the −1 position, since Trp at the −1 position of *Pho*RadA intein could be introduced without significant loss of function. Our results suggest that it might be possible to create "super" inteins that are less extein dependent and highly efficient in protein splicing. Such inteins may resemble an ancient intein that had spread in different host proteins and in various insertion sites. Further atomic-detail structural information for various inteins would be essential for developing promiscuous and robust inteins for advancing protein ligation technology with protein splicing.

Materials and Methods

Protein production of ^{13}C, ^{15}N-labeled *Pho*RadA

Cloning, expression, and purification of *Pho*RadA intein for protein crystallization and X-ray structure determination have been described previously.[28] The same construct of *Pho*RadA intein was also used for preparation of the stable isotope-labeled samples for NMR studies. *Pho*RadA intein was expressed in *E. coli* ER2566 strain where the plasmid (pKRRS15) bearing *Pho*RadA intein was co-transformed with pLysSRARE plasmid to suppress the leaky protein expression and to compensate for rare codons (Novagen). The cells were initially grown in LB medium containing 0.05% (w/v) D-glucose, 100 μg/mL ampicillin, and 5 μg/mL chloramphenicol until the optical density reached $OD_{600} = \sim 0.5$. The cells were collected by spinning down at 900*g* at 25 °C for 10 min, resuspended with M9 medium supplied with 1.3 g/L $^{15}NH_4Cl$ and 2 g/L of 20% or 100% D-[$^{13}C_6$] glucose as sole nitrogen and carbon sources, and induced with a final concentration of 1 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG). ^{13}C, ^{15}N-labeled *Pho*RadA intein was purified as described previously.[28] The buffer was exchanged to 20 mM sodium phosphate buffer (pH 6.0), and protein was concentrated to 0.4 mM in 250 μL. The sample was transferred to a SHIGEMI micro-cell tube for NMR measurements.

Cloning, expression, and purification of $RadA_{min}$ intein

The *cis*-splicing precursor bearing $RadA_{min}$ intein was constructed using the full-length *cis*-splicing *Pho*RadA intein vector (pHYDuet183) as a template.[21] Two oligonucleotides, HK640: 5′-GAT GTT GTT GGT GTT AGG AAT GGA GAA CTT GAG TTC CAT GAG GTT TC and HK641: 5′-GAA ACC TCA TGG AAC TCA AGT TCT CCA TTC CTA ACA CCA ACA ACA TC, were used to remove residues 121–130 of *Pho*RadA intein by PCR, resulting in pDPDuet04.

For structural analysis, inactive $RadA_{min}$ intein was produced as yeast Smt3 fusion protein with a mutation of Cys1 to Ala. In addition, the first residue of C-terminal extein (Thr+1) was replaced by Ala to inactivate the protein splicing completely (Fig. 2). The gene was amplified from pDPDuet04 with two oligonucleotides, HK409: 5′-GCA TAT GGC CTT TGC TAG GGA TAC C and HK390: 5′-GCA TAT GGC CTT TGC TAG GGA TAC C. The PCR product was digested with NdeI and HindIII and cloned into pHYRSF53LA to product a fusion protein, resulting in plasmid pCARSF15.[42] The DNA sequence of pCARSF15 had unexpected insertions of "CAG" at the front of NdeI site, presumably because the 5′ end was not digested by NdeI but directly ligated into the digested vector, possibly due to exonuclease contamination during the digestion of the vector. We decided to retain the insertion because it introduced only an insertion of Gln. The final construct contains a four-residue sequence SQHM as N-extein at the front of *Pho*RadA intein.

E. coli ER2566 strain was transformed with pCARSF15 and grown at 37 °C in LB medium supplied with 25 μg/mL kanamycin. The cells were grown until the optical density reached $OD_{600} = \sim 0.6$, followed by induction with a final concentration of 1 mM IPTG. After 3 h of the induction, the cells were harvested by centrifugation at 6700*g* at 4 °C for 10 min. The cell pellet was resuspended in buffer A [50 mM sodium phosphate buffer and 300 mM NaCl (pH 8.0)] and subsequently flash-frozen in liquid nitrogen for further purification.

The cell suspension was thawed and lysed by ultrasonication. The cell lysate was cleared by centrifugation at 42,000*g* at 4 °C for 45 min. The supernatant was loaded on a 5-mL HisTrap FF column (GE Healthcare). The column was further washed with buffer A and eluted by applying a linear gradient of buffer B [50 mM sodium phosphate, 300 mM NaCl (pH 8.0), and 250 mM imidazole]. Fractions containing $RadA_{min}$ intein fusion protein were dialyzed against phosphate-buffered saline buffer. The dialyzed protein was digested with His-tagged yeast ubiquitin-like-specific protease 1 at 25 °C for 5 h and reloaded on a 5-mL HisTrap FF column (GE Healthcare) to remove digested

His-tagged Smt3 tag and ubiquitin-like-specific protease 1. $RadA_{min}$ intein collected from the flow-through fractions was concentrated by a centrifugal concentrator (Amicon Ultra 3000-molecular-weight cutoff). The buffer was replaced by Milli-Q-grade water.

NMR spectroscopy and resonance assignment

All spectra used for NMR structure determination were recorded at 308 K on either 600- or 800-MHz Varian Innova spectrometers equipped with triple-resonance cryogenic probe heads. The protein was concentrated to 0.15–0.4 mM in 20 mM phosphate buffer (pH 6.0). For sequence-specific resonance assignments, series of two-dimensional and three-dimensional spectra were recorded including [^{1}H,^{15}N]-heteronuclear single quantum correlation (HSQC), [^{1}H,^{13}C]-HSQC, HNCA, intraHNCA, HN(CO)CA, HNCO, HN(CA)CO, HNCACB, CBCA(CO)NH, CCC(CO)NH, HBHA(CO)NH, HBNH, ^{15}N-edited total correlated spectroscopy (TOCSY) with a mixing time of 50 ms, H(C)CH-TOCSY, and H(C)CH-COSY. Aromatic side chain was assigned using (HB)CB(CG)CDHD, (HB)CB(CGCD)CEHE, and aromatic constant time [^{1}H,^{13}C]-HSQC spectra. For the structure calculation, an ^{15}N-edited NOE spectroscopy (NOESY) with a mixing time of 80 ms and a sensitive enhanced ^{13}C-edited NOESY with a mixing time of 100 ms were recorded at ^{1}H frequency of 800 MHz. ^{15}N relaxation times were recorded at ^{1}H frequency of 600 MHz. T_1(^{15}N) relaxation rates were determined using the following T_1 delays: 10, 30, 50, 70, 90, 110, 130, and 150 ms.[43] For T_2(^{15}N) relaxation rates, the following T_2 delays were used: 10, 30, 50, 70, 90, and 110 ms with CPMG refocusing interval of 1.3 ms. Heteronuclear ^{15}N{^{1}H}NOEs were determined by recording [^{1}H,^{15}N]-HSQC spectra with and without 4.0 s of ^{1}H saturation.[43] Peak volumes were used to determine relaxation rates by fitting and analysis in CcpNmr analysis.[44]

All spectra were processed by program NMRPipe, and specific resonance assignments were performed using the software SPARKY.[45] The sequence-specific resonance assignment of the 174-residue *Pho*RadA intein with two C-extein residues was completed by conventional triple-resonance methods using ^{13}C, ^{15}N doubly labeled samples with a final 94% assignment completion of H^N, N, C^α, H^α, and C′ atoms. The aliphatic carbon and proton side-chain resonances are 97% assigned. Proline residue conformation was checked, and they are all in *trans*-conformation based upon C^β and C^γ chemical shift.[46] Residues 121–122, 125–127, 159–160, and 173–174 could not be observed in the [^{1}H,^{15}N]-HSQC spectra (Fig. 1), and resonance assignments of these residues are incomplete.

NMR structure determination

NOESY peak lists were generated with automatic pick picking and integrated using CcpNmr analysis.[44] Backbone-derived angle constraints from TALOS+ were used as inputs for the structure calculation.[25] A chemical shift list, dihedral angle constraints, and unassigned NOESY peak lists were used as inputs for the structure determination using the automatic NOESY assignment approach with CYANA 3.0.[24,47] The structure determination in CYANA was based on the automatic NOE assignment, and in the final cycle of structure determination, 3515 NOEs were used (Table 1). The 20 final solution structures of *Pho*RadA intein with the lowest CYANA target function were further energy refined using AMBER. Molecular dynamic simulation was performed in explicit water with a 5-Å water shell surrounding the protein molecule.[48] The structural statistic and inputs for NMR structure are summarized in Table 1. Structure validation was performed using CING server† and PROCHECK-NMR.[26,49]

Table 3. Crystallographic data collection and structure refinement

	RadA	$RadA_{min}$
Data collection		
Beamline	ESRF/ID23-1	ESRF/ID14-1
Space group	$P2_12_12_1$	$P2_12_12_1$
Molecules per asymmetric unit	2	1
Unit cell parameters		
a, *b*, *c* (Å)	58.1, 67.4, 82.9	46.9, 63.6, 66.6
α, β, γ (°)	90, 90, 90	90, 90, 90
Resolution (Å)[a]	33.7–1.75 (1.80–1.75)	46.0–1.58 (1.64–1.58)
R_{merge} (%)[b]	7.0 (95.3)	8.6 (45.4)
No. of reflections (measured/unique)	234,686/33,297	238,001/27,745
$\langle I/\sigma I\rangle$	17.3 (2.5)	12.2 (2.4)
Completeness (%)	99.3 (99.8)	98.9 (89.6)
Redundancy	7.1 (7.2)	8.6 (7.9)
Refinement		
Resolution (Å)	33.7–1.75	46.0–1.58
No. of reflections (refinement/R_{free})	33,296/999	27,680/887
R/R_{free}[c]	0.190/0.234	0.167/0.195
No. of atoms		
Protein	2788	1366
Ligands	39	0
Water	121	251
RMSD values from ideal		
Bond lengths (Å)	0.012	0.011
Bond angles (°)	1.65	1.4
PDB code	4E2T	4E2U

[a] The values for the highest-resolution shell are shown in parentheses.

[b] $R_{merge} = \sum_h \sum_i |I_i - \langle I\rangle| / \sum_h \sum_i I_i$, where I_i is the observed intensity of the *i*th measurement of reflection *h*, and $\langle I\rangle$ is the average intensity of that reflection obtained from multiple observations.

[c] $R = \sum ||F_o| - |F_c|| / \sum |F_o|$, where F_o and F_c are the observed and calculated structure factors, respectively, calculated for all data. R_{free} was defined in Ref. 51.

Protein crystallization

Crystallization of *Pho*RadA intein has been described previously.[28] $RadA_{min}$ intein was crystallized using solution of 31 mg/mL concentration. The crystallization conditions were screened using solution from Index HT screen (Hampton Research) by sitting-drop vapor diffusion in Innovadyne SD2 96-well plates with drops of

† http://nmr.cmbi.ru.nl/cing/Home.html

100 nL protein solution and 100 nL reservoir solution at 293 K. A single crystal was picked from a drop containing 1.6 M tri-sodium citrate and cryo-protected with Paratone-N before vitrification.

Diffraction data collection and processing

Diffraction data for the crystal of *Pho*RadA intein were collected in a single pass on beamline ID23-1 at European Synchrotron Radiation Facility (ESRF)/Grenoble and were subsequently indexed, integrated, and scaled to 1.75 Å resolution using the program XDS.[50] The crystal belongs to the space group $P2_12_12_1$ with two molecules in the asymmetric unit. Data processing statistics are listed in Table 3. The data set was essentially complete, with high redundancy. However, the values of scaling *R*-factors are rather high, indicating that data quality might not be optimal, but still within an acceptable range. The estimated Matthews coefficient was 1.97 $Å^3$ Da^{-1}, corresponding to 37% solvent content.[52] The structure was solved by molecular replacement using the NMR-derived coordinates as a starting model, with the weak solutions improved with the help of the program Rosetta.[29] Further refinement was performed with Refmac5[53] and PHENIX,[54] using all data between 33.7 and 1.75 Å, after setting aside 3% of randomly selected reflections (~1000 total) for calculation of R_{free}.[51] Non-crystallographic symmetry restraints were initially applied to the two molecules, but these restraints were removed before the final refinement cycles. Isotropic individual temperature factors were refined, with the TLS parameters (one for each molecule) added in the final stages of refinement. After several further rounds of restrained and TLS refinement and manual correction using Coot,[55] the structural model was finally refined to an *R*-factor of 19.0% and an R_{free} of 23.4%. Data processing and structure refinement statistics are shown in Table 3.

Diffraction data for the crystal of $RadA_{min}$ intein were collected on beamline ID14-1 at ESRF/Grenoble. They were indexed, integrated, and scaled to 1.58 Å resolution using the program package HKL2000.[56] A comparatively high value of R_{sym} is due to the presence of a second crystal lattice that was partially superimposed on the main one, interfering with proper integration of a fraction of reflections. This crystal is also in space group $P2_12_12_1$ but with only one molecule in the asymmetric unit. The structure was solved by molecular replacement with the program Phaser using the coordinates of molecule A of *Pho*RadA intein[57] as a search model. The refinement of $RadA_{min}$ used a protocol similar to the one utilized for *Pho*RadA intein, and the final *R*-factor and R_{free} are 16.7% and 19.5%, respectively. Data processing and structure refinement statistics are shown in Table 3.

Mutagenesis of *Pho*RadA intein

Plasmids for the 20 variants of *cis*-splicing *Pho*RadA intein precursor bearing the mutations at the −1 position were constructed from pHYDuet183 by cassette mutagenesis using *Bse*RI and *Hind*III restriction sites or QuikChange Protocol (Stratagene) with synthetic oligonucleotides containing the mutations (the details will be reported elsewhere).[21] The E71T mutation was introduced in the *Pho*RadA intein of the *cis*-splicing precursors containing Asp (pSCFDuet22), Glu (pSCFDuet23), and Lys (pSCFDuet4) at the −1 position. The mutation of E71T was introduced into these plasmids by QuikChange Protocol with the two oligonucleotides, HK976: 5′-CAT ACA TCT ATC GCA CGA AGG TTG AGA AGC and HK977: 5′-GCT TCT CAA CCT TCG TGC GAT AGA TGT ATG.

Evaluation of *cis*-splicing

Cis-splicing efficiency was quantified by transforming the plasmids for the 20 variants at the −1 position into *E. coli* ER2566 cells. The cells were grown in 5 mL LB media supplied with 25 μg/mL kanamycin at 37 °C until OD_{600} reached 0.5–0.6. The expression of the *cis*-splicing precursors was induced by addition of IPTG at a final concentration of 1 mM. After a 4-h induction with IPTG, the cells were harvested by centrifugation at 4 °C for further purification. The cells were lysed by resuspending the pellets in 100 μL B-PER® Bacterial Protein Extraction Reagent (Thermo Scientific) and incubation at 25 °C for 10 min. The cell suspension was cleared by centrifugation at 14,000*g* for 5 min, and the supernatant was loaded on a Ni-NTA spin column (Qiagen). Bound proteins were eluded from the column by applying a buffer containing 50 mM sodium phosphate (pH 8.0), 300 mM NaCl, and 250 mM imidazole. The elution fractions were analyzed on 18% SDS-PAGE (Supplemental Fig. 2). The gels were stained with PhastGel™ Blue R (GE Healthcare), and protein band intensity was determined using the software ImageJ 1.45. The intensities of protein bands were used to quantify the protein splicing efficiency, given that staining dye equally binds to all proteins. The errors for the splicing efficiency were estimated from three independent experiments.

Accession numbers

The resonance assignment is deposited in the BioMagResBank (accession number: 18320). Coordinates and structure factors have been deposited in the PDB with accession numbers 4E2T for *Pho*RadA intein, 4E2U for $RadA_{min}$ intein, and 2LQM for the NMR structures of *Pho*RadA intein.

Acknowledgements

We thank C. Albert and S. Ferkau for technical help in the protein and plasmid preparations. J.S.O. acknowledges the National Graduate School in Informational and Structural Biology for financial support. The WeNMR project [European FP7 e-Infrastructure grant‡, contract no. 261572; supported by the national Grid Initiatives of Belgium, Italy, Germany, the Netherlands (via the Dutch BiG Grid project), Portugal, UK, South Africa, Taiwan,

‡ www.wenmr.eu

and the Latin America Grid infrastructure via the Gisela project] is acknowledged for the use of web portals, computing, and storage facilities. This work was supported in part by the Academy of Finland (1131413 and 137995) and Biocenter Finland (for H.I., the crystallization, and NMR facilities at the Institute of Biotechnology) and in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmb.2012.04.029

References

1. Paulus, H. (2000). Protein splicing and related forms of protein autoprocessing. *Annu. Rev. Biochem.* **69**, 447–496.
2. Perler, F. B. (2002). InBase: the Intein Database. *Nucleic Acids Res.* **30**, 383–384.
3. Pietrokovski, S. (2001). Intein spread and extinction in evolution. *Trends Genet.* **17**, 465–472.
4. Skrisovska, L., Schubert, M. & Allain, F. H. T. (2010). Recent advances in segmental isotope labeling of proteins: NMR applications to large proteins and glycoproteins. *J. Biomol. NMR*, **46**, 51–65.
5. Saleh, L. & Perler, F. B. (2006). Protein splicing *in cis* and *in trans*. *Chem. Rec.* **6**, 183–193.
6. Evans, T. C., Martin, D., Kolly, R., Panne, D., Sun, L., Ghosh, I. *et al.* (2000). Protein *trans*-splicing and cyclization by a naturally split intein from the *dnaE* gene of *Synechocystis* species PCC6803. *J. Biol. Chem.* **275**, 9091–9094.
7. Iwai, H., Lingel, A. & Plückthun, A. (2001). Cyclic green fluorescent protein produced *in vivo* using an artificially split PI-*PfuI* intein from *Pyrococcus furiosus*. *J. Biol. Chem.* **276**, 16548–16554.
8. Iwai, H., Züger, S., Jin, J. & Tam, P. (2006). Highly efficient protein *trans*-splicing by a naturally split DnaE intein from *Nostoc punctiforme*. *FEBS Lett.* **580**, 1853–1858.
9. Lockless, S. W. & Muir, T. W. (2009). Traceless protein splicing utilizing evolved split inteins. *Proc. Natl Acad. Sci. USA*, **106**, 10999–11004.
10. Noren, C. J., Wang, J. & Perler, F. B. (2000). Dissecting the chemistry of protein splicing and its applications. *Angew. Chem., Int. Ed. Engl.* **39**, 450–466.
11. Southworth, M. W., Benner, J. & Perler, F. B. (2000). An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile. *EMBO J.* **19**, 5019–5026.
12. Tori, K., Dassa, B., Johnson, M. A., Southworth, M. W., Brace, L. E., Ishino, Y. *et al.* (2010). Splicing of the mycobacteriophage Bethlehem DnaB intein: identification of a new mechanistic class of inteins that contain an obligate block F nucleophile. *J. Biol. Chem.* **285**, 2515–2526.
13. Klabunde, T., Sharma, S., Telenti, A., Jacobs, W. R. & Sacchettini, J. C. (1998). Crystal structure of GyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing. *Nat. Struct. Biol.* **5**, 31–36.
14. 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.
15. Sun, P., Ye, S., Ferrandon, S., Evans, T. C., Xu, M. Q. & 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.
16. Johnson, M. A., Southworth, M. W., Herrmann, T., Brace, L., Perler, F. B. & Wüthrich, K. (2007). NMR structure of a KlbA intein precursor from *Methanococcus jannaschii*. *Protein Sci.* **16**, 1316–1328.
17. Amitai, G., Callahan, B. P., Stanger, M. J., Belfort, G. & Belfort, M. (2009). Modulation of intein activity by its neighboring extein substrates. *Proc. Natl Acad. Sci. USA*, **106**, 11005–11010.
18. Nogami, S., Satow, Y., Ohya, Y. & Anraku, Y. (1997). Probing novel elements for protein splicing in the yeast Vma1 protozyme: a study of replacement mutagenesis and intragenic suppression. *Genetics*, **147**, 73–85.
19. Aranko, A. S., 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.
20. Chong, S., Shao, Y., Paulus, H., Benner, J., Perler, F. B. & Xu, M. Q. (1996). Protein splicing involving the *Saccharomyces cerevisiae* VMA intein. The steps in the splicing pathway, side reactions leading to protein cleavage, and establishment of an *in vitro* splicing system. *J. Biol. Chem.* **271**, 22159–22168.
21. Ellilä, S., Jurvansuu, J. M. & Iwaï, H. (2011). Evaluation and comparison of protein splicing by exogenous inteins with foreign exteins in *Escherichia coli*. *FEBS Lett.* **585**, 3471–3477.
22. Busche, A. E. L., Aranko, A. S., Talebzadeh-Farooji, M., Bernhard, F., Dötsch, V. & Iwaï, H. (2009). Segmental isotopic labeling of a central domain in a multidomain protein by protein *trans*-splicing using only one robust DnaE intein. *Angew. Chem., Int. Ed. Engl.* **48**, 6128–6131.
23. Shah, N. H., Vila-Perelló, M. & 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.
24. Güntert, P. (2009). Automated structure determination from NMR spectra. *Eur. Biophys. J.* **38**, 129–143.
25. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. (2009). TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, **44**, 213–223.
26. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.
27. Perler, F. B. (1998). Protein splicing of inteins and hedgehog autoproteolysis: structure, function, and evolution. *Cell*, **92**, 1–4.

28. Lyskowski, A., Oeemig, J. S., Jaakkonen, A., Rommi, K., DiMaio, F., Zhou, D. *et al.* (2011). Cloning, expression, purification, crystallization and preliminary X-ray diffraction data of the *Pyrococcus horikoshii* RadA intein. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **67**, 623–626.
29. DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U. *et al.* (2011). Increasing the radius of convergence of molecular replacement by density and energy guided protein structure optimization. *Nature*, **473**, 540–543.
30. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291.
31. Jaudzems, K., Geralt, M., Serrano, P., Mohanty, B., Horst, R., Pedrini, B. *et al.* (2010). NMR structure of the protein NP_247299.1: comparison with the crystal structure. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **66**, 1367–1380.
32. Oeemig, J. S., Aranko, A. S., Djupsjöbacka, J., Heinämäki, K. & Iwai, 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.
33. Mizutani, R., Nogami, S., Kawasaki, M., Ohya, Y., Anraku, Y. & Satow, Y. (2002). Protein-splicing reaction *via* a thiazolidine intermediate: crystal structure of the *VMA1*-derived endonuclease bearing the N and C-terminal propeptides. *J. Mol. Biol.* **316**, 919–929.
34. Ludwig, C., Schwarzer, D. & Mootz, H. D. (2008). Interaction studies and alanine scanning analysis of a semi-synthetic split intein reveal thiazoline ring formation from an intermediate of the protein splicing reaction. *J. Biol. Chem.* **283**, 25264–25272.
35. Poland, B. W., Xu, M. Q. & Quiocho, F. A. (2000). Structural insights into the protein splicing mechanism of PI-SceI. *J. Biol. Chem.* **275**, 16408–16413.
36. 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.
37. Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*, 2nd edit. W. H. Freeman and Company, New York, NY.
38. Hackeng, T. M., Griffin, J. H. & Dawson, P. E. (1999). Protein synthesis by native chemical ligation: expanded scope by using straightforward methodology. *Proc. Natl Acad. Sci. USA*, **96**, 10068–10073.
39. Southworth, M. W., Amaya, K., Evans, T. C., Xu, M. Q., and Perler, F. B. (1999). Purification of proteins fused to either the amino or carboxy terminus of the *Mycobacterium xenopi* gyrase A intein. BioTechniques 27, 110–114, 116, 118–120.
40. Chong, S., Montello, G. E., Zhang, A., Cantor, E. J., Liao, W., Xu, M. Q. & Benner, J. (1998). Utilizing the C-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step. *Nucleic Acids Res.* **26**, 5109–5115.
41. Inglis, A. S. (1983). Cleavage at aspartic acid. *Methods Enzymol.* **91**, 324–332.
42. Muona, M., Aranko, A. S. & Iwai, H. (2008). Segmental isotopic labelling of a multidomain protein by protein ligation by protein trans-splicing. *ChemBioChem*, **9**, 2958–2961.
43. Farrow, N. A., Muhandiram, R., Singer, A. U., Pascal, S. M., Kay, C. M., Gish, G. *et al.* (1994). Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry*, **33**, 5984–6003.
44. Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M. *et al.* (2005). The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, **59**, 687–696.
45. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.
46. Schubert, M., Labudde, D., Oschkinat, H. & Schmieder, P. (2002). A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on ^{13}C chemical shift statistics. *J. Biomol. NMR*, **24**, 149–154.
47. Herrmann, T., Güntert, P. & Wüthrich, K. (2002). Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319**, 209–227.
48. Bertini, I., Case, D. A., Ferella, L., Giachetti, A. & Rosato, A. (2011). A Grid-enabled web portal for NMR structure refinement with AMBER. *Bioinformatics*, **27**, 2384–2390.
49. Doreleijers, J. F., Vranken, W. F., Schulte, C. S., Markley, J. L., Ulrich, E. L., Vriend, G. & Vuister, G. W. (2012). NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. *Nucleic Acids Res.* **40**, D519–D524.
50. Kabsch, W. (1993). Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr.* **26**, 795–800.
51. Brünger, A. T. (1992). The free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**, 472–475.
52. Matthews, B. W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497.
53. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **53**, 240–255.
54. Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W. *et al.* (2002). PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **58**, 1948–1954.
55. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **60**, 2126–2132.
56. Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326.
57. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674.