Protein *trans*-splicing and its use in structural biology: opportunities and limitations

Gerrit Volkmann and Hideo Iwaï*

Received 6th June 2010, Accepted 21st July 2010
DOI: 10.1039/c0mb00034e

Obtaining insights into the molecular structure and dynamics of a protein by NMR spectroscopy and other in-solution biophysical methods relies heavily on the incorporation of isotopic labels or other chemical modifications such as fluorescent groups into the protein of interest. These types of modifications can be elegantly achieved with the use of split inteins in a site- and/or region-specific manner. Split inteins are split derivatives of the protein splicing element intein, and catalyze the formation of a peptide bond between two proteins. Recent progress in split intein engineering provided the opportunity to also perform peptide bond formation between a protein and a chemically synthesized peptide. We review the current state-of-the-art in preparing segmental isotope-labeled proteins for NMR spectroscopy, and highlight the importance of split intein orthogonality for the ligation of a protein from multiple fragments. Furthermore, we use split intein-mediated site-specific fluorescent labeling as a framework to illustrate the general usefulness of split inteins for custom protein modifications in the realm of structural biology. We also address some limitations of split intein technology, and offer constructive advice to overcome these shortcomings.

1. Introduction

Structural biology is engaged in analyzing ever larger and more complex systems, driven by the fact that a full understanding of biological function requires to target the multi-protein assembly in which a biomolecule is constitutively or transiently involved. Quantitative analysis of protein complexes, protein structures and their dynamics *in situ* could shed lights on how proteins actually function in living organisms through structural changes and interactions with other biomolecules. Modern biophysical methods such as X-ray crystallography can provide high-resolution three-dimensional structures of large multi-component systems of even over a few megadalton, although it requires crystallization and largely lacks dynamical aspects of the systems. Optical methods including Fluorescence Resonance Energy Transfer (FRET) and Electron Spin Resonance (ESR) spectroscopy have been used to detect molecular assemblies and structural changes of proteins at low resolution, which can be applied to any size of systems with information on dynamics ranging from picoseconds to minutes. Nuclear Magnetic Resonance (NMR) spectroscopy is unique in providing high-resolution three-dimensional protein structures, their populations, and dynamics ranging from picoseconds to days, despite the molecular size limit. Recent

Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki, Helsinki, Finland. E-mail: hideo.iwai@helsinki.fi

Gerrit Volkmann

Gerrit Volkmann (born in 1981, Lüneburg, Germany) obtained his MSc in biochemistry from University of Bremen, Germany, in 2006. A fellowship from the Deutscher Akademischer Auslandsdienst allowed him to conduct his MSc thesis in the laboratory of Dr Xiang-Qin Liu at Dalhousie University, Halifax, Canada, where he continued his doctoral studies on intein biochemistry and protein engineering. After obtaining his PhD in 2009, he joined the group of Dr Hideo Iwai at the University of Helsinki, Finland, to explore segmental isotope labeling of pharmaceutically relevant proteins.

Hideo Iwai

Hideo Iwai obtained Bsc and MSc (Pharmaceutical Science) from the University of Tokyo, Japan. He completed his Dr sc. nat. (Biology) at the Institute of Molecular Biology & Biophysics, ETH-Zürich, Switzerland. He is currently a group leader at the Institute of Biotechnology, University of Helsinki, Finland, and Academy Research Fellow since 2006. His current research focuses on the molecular mechanism of protein splicing at atomic resolution and the development of intein-based protein engineering technology for NMR spectroscopy.

progresses in NMR techniques and instruments have extended the size limitation, and it is now possible to study biological macromolecules or supra-molecular assemblies with masses higher than 100 kDa.[1] However, structural analysis by biophysical methods like optical, ESR, or NMR spectroscopy has often been hampered by available probes (*e.g.* isotopes, electrospins and fluorophores). The importance of labeled samples has become apparent, *e.g.* by the fact that structure determination of proteins up to 20–30 kDa by NMR spectroscopy has tremendously advanced since preparation of ^{13}C, ^{15}N doubly isotope-labeled samples from *E. coli* became a routine procedure. Advanced selective isotopic labeling of proteins has extended applications of NMR spectroscopy to even bigger proteins. ^{1}H,^{13}C methyl-labeled, highly deuterated proteins in concert with experiments that exploit the methyl-TROSY effect have facilitated the study of very high molecular weight proteins of >200 kDa although site-specific assignments still remain the main challenge.[2] The application of optical approaches such as FRET is limited to proteins containing two or more fluorescent probes, even though there is no size limitation. Selectively labeled proteins with fluorophores could also widely be used for analysis of intermediate protein structures during protein folding, and dynamics of interactions with biomolecules. Thus, the limiting step in structural studies of proteins and protein complexes *in vitro* as well as *in vivo* by biophysical methods is frequently the preparation of proteins of interests with desired labels at specific sites and/or regions. Because structural analysis of transiently formed complexes and structural studies in living systems are very challenging with conventional approaches, site-specifically labeled samples could play a vital role in extracting structural information from these complex systems. Protein *trans*-splicing (PTS) has great potential to become an indispensable tool that could advance the current applications of biophysical methods in studying intact multi-domain proteins and larger protein complexes by facilitating site- and/or region-specific labeling of proteins *in vitro* as well as *in vivo*.

2. Protein *trans*-splicing: split inteins

Protein *trans*-splicing is a remarkable biological process, whereby a full-length protein is reconstituted from two fragments through the formation of a peptide bond[3,4] (Fig. 1b). This protein ligation reaction is catalyzed by split inteins, which represent a distinct subgroup of protein splicing elements referred to as inteins[5] (Fig. 1a). The term intein is derived from *in*ternal pro*tein* because inteins are imbedded into the open reading frame of a host protein, much like introns in pre-mRNA. Consequently, the flanking host protein sequences are called exteins (from *ex*ternal pro*tein*), and the primary translation product is usually referred to as the precursor protein. In the case of split inteins, the open reading frame of the precursor protein is fragmented into two separate genes,

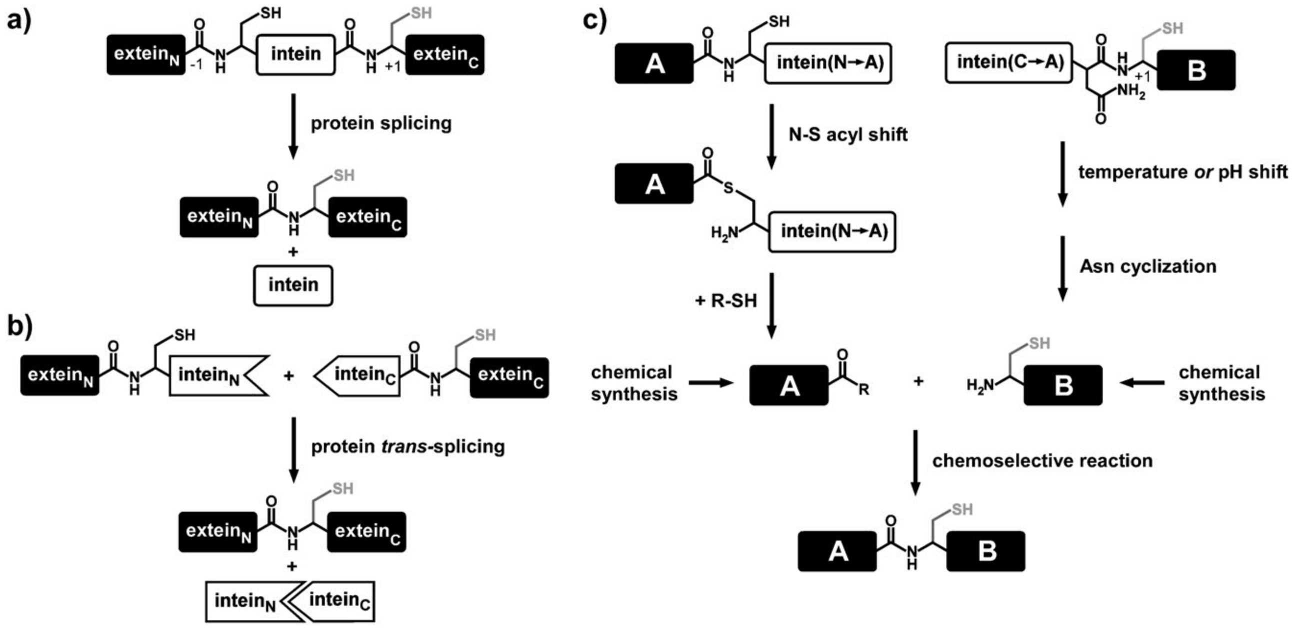

Fig. 1 Protein ligation strategies. Protein ligation by protein splicing utilized by contiguous inteins (a) and split inteins (b). Extein$_N$ and extein$_C$ refer to the N- and C-terminal extein sequences, respectively. The side chain of the nucleophilic residue at position +1 is shaded in grey, and is retained in the ligation product. Both the +1 residue, as well as the nucleophilic residue at the beginning of the intein are shown as cysteines, although they can also be Ser (at position 1) or Ser and Thr (at position +1) in class I inteins. (c) shows a schematic representation of protein ligation between two reactants A and B by expressed protein ligation (EPL) and native chemical ligation (NCL). The term NCL is used when both ligation reactants are generated by chemical synthesis, EPL is used when at least one reactant is prepared by recombinant methods (top half). A C-terminal thioester on a recombinant protein is conveniently prepared using a self-cleavable intein tag bearing a mutation of the terminal Asn residue (N → A). An N-terminal Cys residue is commonly achieved by employing a self-cleavable intein tag carrying a mutation of the nucleophilic residue at position 1 of the intein (*e.g.* C → A) or proteolysis. This Cys residue is thus a prerequisite for protein ligation using both NCL and EPL, and is retained in the ligation product.

with the break in sequence occurring within the intein, thus the designation "split intein".

Protein ligation by protein *trans*-splicing does not require energy equivalents such as ATP but is solely dictated by the intein structure encoded in the primary structure, along with the first residue of the C-terminal extein (called the +1 residue). The currently accepted canonical protein splicing mechanism proceeds in four concerted nucleophilic displacement reactions[6] (Fig. 2), which is preceded by an association step between the N- and C-terminal intein halves in the case of split inteins. In the initial reaction, the N-terminal extein sequence is transferred to the side chain of intein residue 1 by an N–X acyl rearrangement (X denoting either a sulfur (S) or oxygen (O) atom), resulting in a linear ester intermediate. In the second step, *trans*-esterification, the nucleophilic +1 residue of the C-extein attacks the ester bond of the linear intermediate, resulting in the formation of a branched ester intermediate, where the N-extein is linked to the C-extein by an ester bond. The third step is catalyzed by the last intein residue (most often Asn), which cyclizes to form a succinimide ring, whereby the intein (or C-terminal split intein half) is cleaved off the branched intermediate. The last step is a spontaneous X–N acyl rearrangement between the esterified exteins due to energetically favourable peptide formation, resulting in the final formation of the peptide bond and thus the ligated host protein.

2.1 Engineered split inteins

Split inteins can be artificially engineered from contiguous inteins by dividing the intein sequence into two parts. In fact, artificial split inteins were investigated even before the first natural split intein was reported.[3,7,8] Splitting of inteins can be done both within the protein-splicing domain and the endonuclease domain (present only in bi-functional inteins). Artificial split inteins have also been created from natural split inteins by first fusing the two split intein halves together and then introducing the split site at a location different from the natural site.[9,10] Naturally occurring split inteins only contain a protein-splicing domain, with the split site corresponding to the location of the endonuclease domain in bi-functional inteins.[4] This location often represents a flexible loop region within the intein structure,[10,11,12] which is able to tolerate the introduction of a split site.

Recent advances in split intein engineering have revealed that split inteins can also be created by introducing split sites at locations differing from the canonical split site of naturally occurring split inteins (Fig. 3), however, the location of the split site in a loop region appears to be a general requirement to retain splicing function. The most intriguing non-canonical split inteins reported are those in which the split site was brought in close proximity to either end of the intein sequence. For example, the *Ssp* DnaB S1 split intein catalyzes protein *trans*-splicing with an N-intein fragment that is 11 aa long,[13] which is much smaller than the ~125 aa long N-intein fragments of naturally occurring split inteins. Along similar lines, split inteins with shortened C-terminal fragments have successfully been engineered from both the naturally occurring *Npu* DnaE split intein and a mini-intein derived from the bi-functional *Ssp* GyrB intein, with C-intein sequences as short as 6 aa,[10,14] a significant reduction in size compared to the ~30 aa long C-intein fragments of natural split inteins, although the ligation efficiencies are generally lower than at the original sites.

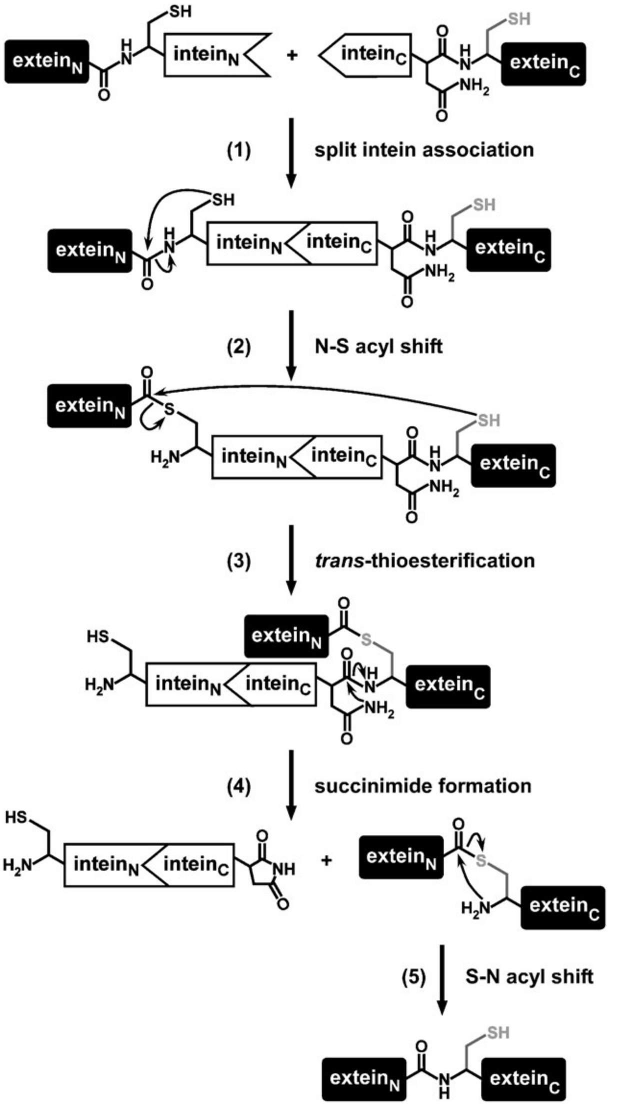

Fig. 2 Canonical protein splicing mechanism utilized by split inteins. The N-precursor protein (comprising extein$_N$ and intein$_N$) and C-precursor protein (containing intein$_C$ and extein$_C$) are expressed from separate genes, followed by association of the intein$_N$ and intein$_C$ parts of the split intein (step 1). The assembled split intein is now active to catalyze the first step in the protein splicing reaction, an N–S acyl shift involving the first residue of intein$_N$ (step 2). The thioester intermediate is then attacked by the first residue of extein$_C$ in a *trans*-thioesterification reaction (step 3), which also leads to physical separation of intein$_N$ from extein$_N$. Next, the last residue of intein$_C$ (Asn) forms a succinimide ring, effectively cleaving intein$_C$ from the esterified exteins (branched intermediate) (step 4). The split intein likely remains assembled after the *trans*-splicing reaction. The final reaction is a spontaneous S–N acyl shift between the esterified exteins, leading to peptide bond formation between extein$_N$ and extein$_C$ (step 5). Although the reactions shown in this scheme involve Cys residues at position 1 of both intein$_N$ and extein$_C$, other residues (Ser and Thr) at these positions are possible.

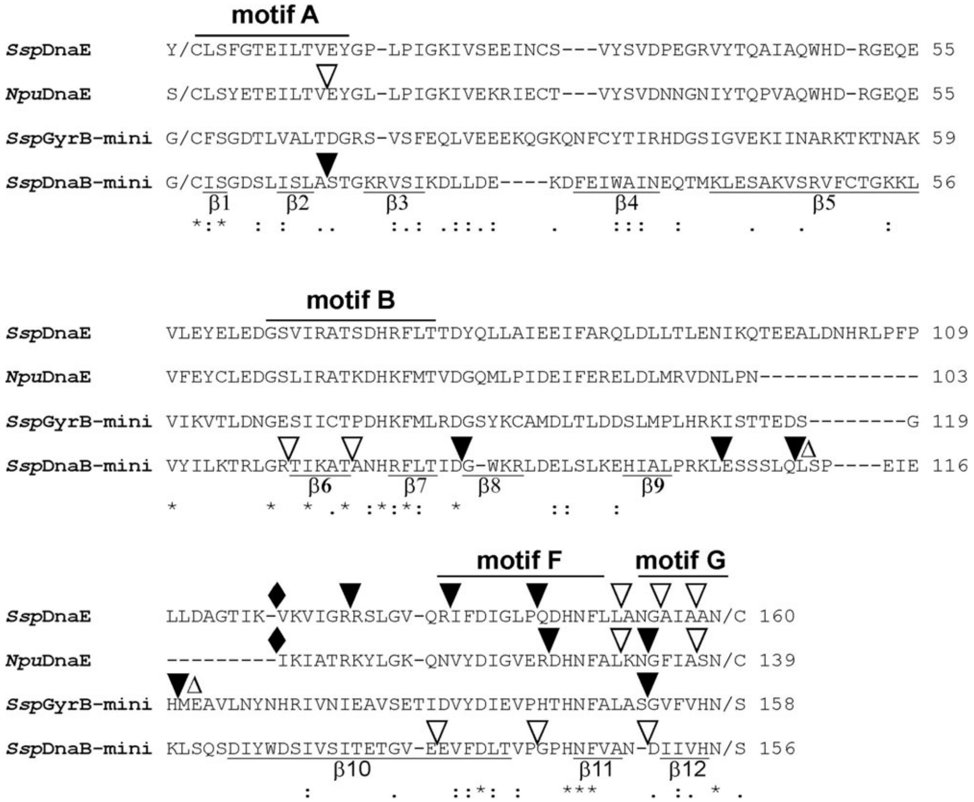
motif A
SspDnaE Y/CLSFGTEILTVEYGP-LPIGKIVSEEINCS---VYSVDPEGRVYTQAIAQWHD-RGEQE 55
NpuDnaE S/CLSYETEILTVEYGL-LPIGKIVEKRIECT---VYSVDNNGNIYTQPVAQWHD-RGEQE 55
SspGyrB-mini G/CFSGDTLVALTDGRS-VSFEQLVEEEKQGKQNFCYTIRHDGSIGVEKIINARKTKTNAK 59
SspDnaB-mini G/CISGDSLISLASTGKRVSIKDLLDE----KDFEIWAINEQTMKLESAKVSRVFCTGKKL 56
β1
β2
β3
β4
β5
: : : .. :.: .::.: . ::: : . . :
motif B
SspDnaE VLEYELEDGSVIRATSDHRFLTTDYQLLAIEEIFARQLDLLTLENIKQTEEALDNHRLPFP 109
NpuDnaE VFEYCLEDGSLIRATKDHKFMTVDGQMLPIDEIFERELDLMRVDNLPN------------- 103
SspGyrB-mini VIKVTLDNGESIICTPDHKFMLRDGSYKCAMDLTLDDSLMPLHRKISTTEDS--------G 119
SspDnaB-mini VYILKTRLGRTIKATANHRFLTIDG-WKRLDELSLKEHIALPRKLESSSLQLSP----EIE 116
Δ
β6
β7
β8
β9
* * * .* :*:*: * :: :
motif F
motif G
SspDnaE LLDAGTIK-VKVIGRRSLGV-QRIFDIGLPQDHNFLLANGAIAAN/C 160
NpuDnaE ---------IKIATRKYLGK-QNVYDIGVERDHNFALKNGFIASN/C 139
Δ
SspGyrB-mini HMEAVLNYNHRIVNIEAVSETIDVYDIEVPHTHNFALASGVFVHN/S 158
SspDnaB-mini KLSQSDIYWDSIVSITETGV-EEVFDLTVPGPHNFVAN-DIIVHN/S 156
β10
β11
β12
: . ::*: : *** . :. * .

Fig. 3 Naturally and artificially split intein sequences. The sequences of the natural *Ssp* and *Npu* DnaE split inteins and the engineered *Ssp* GyrB and DnaB mini-inteins were aligned using ClustalW. Intein sequence motifs are given above the sequences. Also indicated below the *Ssp* DnaB sequence are the β-strands found in the *Ssp* DnaB crystal structure (β1 to β12). Black arrowheads mark the introduction of non-canonical split sites into an intein without losing protein-splicing function. The Δ symbol next to a black arrowhead indicates the deletion of an intein endonuclease sequence at this position. White arrowheads correspond to non-canonical split site insertions that resulted in non-functional split inteins. Filled black diamonds indicate the split site of the two natural split inteins.

Split inteins with extremely short N- and C-terminal halves are of significant interest for protein engineering as 'ligation tags', because they can provide a facile means for the site-specific incorporation of unnatural amino acids, fluorescent labels or other biophysical probes into a protein in combination with chemical synthesis. The N- or C-intein can be produced by standard solid-phase peptide synthesis, and the label is integrated into a short extein sequence. Protein *trans*-splicing between the peptide and a recombinant protein containing the remainder of the non-canonical split intein fused to the target protein then results in appendage of the 'labeled extein' to the target protein.[15,16] The advantage of using these non-canonical split intein fragments for protein modification and labeling instead of the natural split inteins is clearly their reduced size, making their chemical synthesis less laborious and much more cost-effective. It is not without reason why, for example, the 36 aa *Ssp* DnaE C-intein was so rarely exploited for protein modification since its discovery in 1998.[17–19]

2.2 Application of split inteins

Apart from protein labeling, which can be accomplished using either the semi-synthetic strategies described or purely recombinant approaches,[20–22] split inteins have become powerful tools in other protein engineering areas. Due to the general promiscuity of split inteins towards extein sequences, one can perform a multitude of protein ligations depending on the purpose of the investigation. (1) Modular proteins can be assembled from precursor fragments if production of the full-length protein is problematic. This has for example been applied to generate full-length phosphoinositide-dependent kinase (PDK) 1 from precursors containing the N-terminal catalytic kinase domain and the C-terminal pleckstrin homology (PH)-domain,[23] as well as the signal adaptor protein c-CrkII.[24] (2) Single domain globular proteins can be reconstituted from two fragments into a biologically active full-length protein. This reconstitution approach has been the basis for cell-based assays to monitor protein–protein interactions[25] and the spatio-temporal expression of proteins using split reporter proteins, like enhanced green fluorescent protein[26] and luciferase,[27] for the production of environmentally safe transgenic plants by reconstitution of herbicide-resistance proteins from inactive precursor fragments,[28,29] for gene therapy,[30] and recently for reconstitution of a β-barrel membrane protein.[31] (3) Linear polypeptide chains can be cyclized by inserting the desired protein sequence between a circularly permuted split intein, so that as a result of protein *trans*-splicing, the N- and C-termini of the target protein are joined by a peptide bond.[32–36] Alternatively to intein technology, peptide bond formation between two protein partners can also be accomplished by native chemical ligation,[37] which is a

Table 1 Orthogonality of naturally occurring and artificially split inteins. '–' indicates combinations of split intein fragments I_N and I_C, which do not cross-splice. Splicing of endogenous combinations are indicated by '+'. Blanks refer to combinations, which have not been tested for cross-reactivity.

	I_N					
I_C	*Ssp* DnaB	*Sce* VMA	PI-*Pfu*I	PI-*Pfu*II	*Ssp* DnaE	*Mxe* GyrA
Ssp DnaB	+	–[67]				–[21]
Sce VMA	–[67]	+			–[24]	
PI-*Pfu*I			+	–[55]		
PI-*Pfu*II			–[55]	+		
Ssp DnaE		–[24]			+	
Mxe GyrA	–[21]					+

chemoselective reaction between a C-terminal thioester moiety on one partner and an N-terminal Cys residue on the other (Fig. 1c).

2.3 Orthogonality of split inteins

To extend split intein based protein engineering from two-fragment ligation to ligation of three or more fragments, it is necessary to use functionally orthogonal split inteins in order to prevent undesired side products due to cross-reactivity, *e.g.* cyclized proteins. Several naturally occurring and artificially split inteins have been examined for their orthogonality. The natural DnaE split inteins from *Nostoc punctiforme* and *Synechocystis* sp. PCC 6803 cross-splice,[38] as do the DnaE split inteins from three other cyanobacteria[39] (*Nostoc* sp. PCC 7120, *Oscillatoria limnetica* and *Thermosynechococcus vulcanus*). This is not surprising since the DnaE split inteins share a high degree of sequence identity and similarity (52–68% and 72–85%, respectively). Hence, it is reasonable to assume that the *Npu* and *Ssp* DnaE split inteins will also cross-splice with the *Nsp*, *Oli* and *Tvu* DnaE split inteins, although this remains to be confirmed. The few orthogonal split intein combinations reported so far are given in Table 1, however, many combinations have yet to be characterized for their orthogonality to fully explore multi-fragment protein ligation.

Non-canonical split inteins are especially interesting candidates to exploit orthogonality. By shifting the split site up or downstream of the natural site, split intein combinations can be created whose orthogonality is based on extensive sequence overlaps and gaps in the sequence. This was shown for the *Npu* DnaE intein, where the natural split intein could be used in combination with an artificial split intein, which had its split site shifted 21 residues downstream of the natural site, for the assembly of a protein from three fragments without undesired side reactions.[40] The laboratory has since examined other combinations of the natural and artificially split *Npu* and *Ssp* DnaE inteins, with the data presented in Table 2.[9,10,40] These and future combinations will provide the necessary tools to extend protein ligation from more than three fragments.

3. Segmental isotopic labeling

Conventional NMR techniques are generally limited to proteins with molecular weights below 25–30 kDa. Larger proteins or proteins with repetitive sequences or domains not only produce more complex spectra with extensive signal overlap due to an increased number of NMR-active nuclei, but the slower tumbling of the large molecules also shortens transverse spin relaxation, ultimately causing an increase in signal line width and a decrease in sensitivity, thereby making spectral assignment even more challenging.[1,41] One remedy for signal overlap is to reduce the number of signals by incorporating stable isotopes site- or region-specifically. Segmental isotopic labeling, in which only a segment of a protein sequence is labeled, is an ideal approach for NMR analysis, as the protein can still be analyzed by conventional triple-resonance assignment approaches. Fig. 4 shows general strategies to incorporate isotope labels in only a portion of a protein using split intein mediated protein *trans*-splicing (PTS). Native chemical ligation (NCL; also called expressed protein ligation, EPL) has also been exploited for preparing segmental isotope-labeled proteins.[42–48] However, this method requires preparing intermediate thiol products *in vitro* prior to protein ligation. In contrast, protein *trans*-splicing requires no additional thiol reagent or co-factor to ligate two polypeptide chains, permitting protein ligation *in vivo*. Table 3 gives a more detailed comparison of PTS and NCL/EPL.

Table 2 Orthogonality of naturally occurring and artificial DnaE split inteins. '–' indicates split intein fragment combinations, which do not cross-splice. The combinations for functional *trans*-splicing with a model system are indicated by '+'. For the I_C fragments, the subscripts indicate the length of I_C (in amino acid residues) counting from the C-terminal end of the complete intein. The lengths of the I_N fragments are indicated by subscripts, which refer to the complete intein without the indicated intein residues counting from the C-terminal end. For instance, the I_N/I_C pair ΔC36/C36 of *Ssp* DnaE corresponds to the natural split intein, counting Met of the start codon.

	I_C			
I_N	*Ssp* DnaE_{C36}	*Npu* DnaE_{C36}	*Npu* DnaE_{C15}	*Npu* DnaE_{C6}
Ssp $\text{DnaE}\Delta_{\text{C36}}$	+	+		
Ssp $\text{DnaE}\Delta_{\text{C16}}$	+	+		
Npu $\text{DnaE}\Delta_{\text{C36}}$	+	+	–	–
Npu $\text{DnaE}\Delta_{\text{C15}}$	+	+	+	–
Npu $\text{DnaE}\Delta_{\text{C6}}$	+	+	+	+

Segmental isotopic labeling of N- and C-terminal protein segments was first established *in vitro* using purified precursor proteins and an artificially split PI-*Pfu*I intein as the mediator of PTS.[49] Though the target protein in this initial report (C-terminal domain of the RNA polymerase α subunit) was very small (∼9 kDa), this system was later used successfully for preparing segmental isotope-labeled maltose binding protein (42 kDa),[50] and allowed for the near complete resonance assignment of the F_0F_1 ATPase β subunit (52 kDa),[51] which

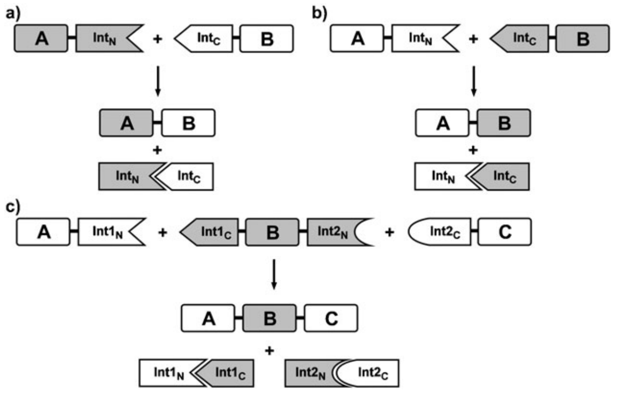

Fig. 4 Strategies for segmental isotope labeling of proteins using split inteins. Shown are schematic illustrations for generating segmental isotope labeled protein labeled (a) in an N-terminal segment, (b) in a C-terminal segment, and (c) in an internal segment using two orthogonal split inteins (c). The grey shading indicates the presence of isotope labels. The examples shown can be generated both *in vitro* and *in vivo* (see text for details). More complex isotopic labeled proteins can also be achieved by *e.g.* preparing two precursor proteins in separate media with different isotope labels. Int$_N$: N-terminal split intein fragment; Int$_C$: C-terminal split intein fragment; A, B, C: segments or modules in a polypeptide chain.

proved that isotopic labeling in defined segments can facilitate structure analysis of proteins larger than 20 kDa by NMR spectroscopy.

Although ground-breaking, the preparation of the segmental isotope-labeled proteins using the PI-*Pfu*I intein *in vitro* was often a time-consuming process, and optimization of the techniques strongly depended on the individual target protein and could thus not be easily transferred from one protein to another. These and other obstacles (see below), together with the only moderate to low yields of final product, might have been the reason why the *in vitro* PTS system for preparation of segmental isotope-labeled protein remained under-appreciated. Advances to make split inteins more attractive for this purpose were made by allowing the ligation step to proceed *in vivo* rather than *in vitro*. The underlying idea of the *in vivo* approach is to express the split intein-containing precursor proteins at different times within a single culture, and to perform an exchange of the growth medium from *e.g.* unlabeled to labeled conditions between the individual expression steps[52,53] (Fig. 5). In this way, only one of the precursor proteins (and thus only one segment of the final target protein) would be isotopically labeled.

The feasibility of this *in vivo* system was first shown for the production of labeled target proteins with unlabeled solubility-enhancing tag proteins. Because some proteins are insufficiently soluble for NMR studies when expressed in recombinant form, the addition of a solubility-enhancement tag (SET) can prevent their aggregation *in vivo*, however, SET itself should be free of isotope labels in order not to interfere with the signals of the target protein during NMR spectroscopy. Using this approach, the prion-inducing domain of yeast Sup35p, which usually forms spontaneous aggregates upon recombinant expression, could be stabilized in a soluble form by ligation to domain B1 of the immunoglobulin binding protein G (GB1) as a SET *in vivo* using the natural *Ssp* DnaE split intein.[54] NMR spectroscopy on the Sup35p–GB1 fusion protein only gave signals for the isotope-labeled part (Sup35p), as anticipated. Isotope scrambling, which refers to the undesired incorporation of isotope labels into the solubility tag due to metabolic flux, could be diminished to negligible levels ($<3\%$) by employing a simple wash step prior to switching to isotope-free medium, thereby affording the clean spectra for the Sup35p protein.[54] Advancement of this *in vivo* segmental isotopic labeling approach also allowed for the preparation of a modular protein labeled either in the N- or C-terminal domain.[52]

3.1 Multi-fragment ligation: segmental isotopic labeling of a central protein fragment using orthogonal split inteins

Two-fragment ligation for segmental isotopic labeling can be only useful when the regions of interest for labeling are located not far from the termini. In larger proteins (>50 kDa), segmental isotopic labeling might be no longer effective when the sites of interest are located in a central part. Similarly, when a protein contains more than two modules of a repeating sequence, two-fragment ligation strategies cannot solve the problem of signal overlap. Segmental isotopic labeling of a central protein fragment provides an ideal solution to this problem. Assembly of a full-length protein from three fragments A, B and C, with only the B fragment containing isotope labels, requires the use of two orthogonal split inteins (Fig. 4c). For successful ligation of the full-length ABC

Table 3 Comparison of reaction specifics between protein *trans*-splicing (PTS) and native chemical ligation (NCL)/expressed protein ligation (EPL)

	PTS	NCL/EPL
Minimal reactant concentrations[a]	nM to µM	mM
Reaction time	min to h	h to days
Amino acid required at the C-terminal junction at ligation point	Cys, Ser, Thr[b]	Cys
N-terminal junction residue	Dependent on inteins	Preferably Gly or Ala[75,c]
Affinity between reactants	Yes, provided by split intein fragments	No
Sensitive to denaturants	Yes/no[d]	No
Additional reagent	No	Yes (thiol reagent)
In vivo ligation	Yes	No
Multi-fragment ligation	One pot/stepwise	Stepwise

[a] To achieve optimal yield. [b] Dependent on intein; adjacent residues might also affect final yield. [c] β-Branched amino acid directly N-terminal to Cys reduces final yield.[37] [d] *Npu* DnaE and *Psp* Pol-1 split inteins splice well in buffer containing up to 6 M urea.

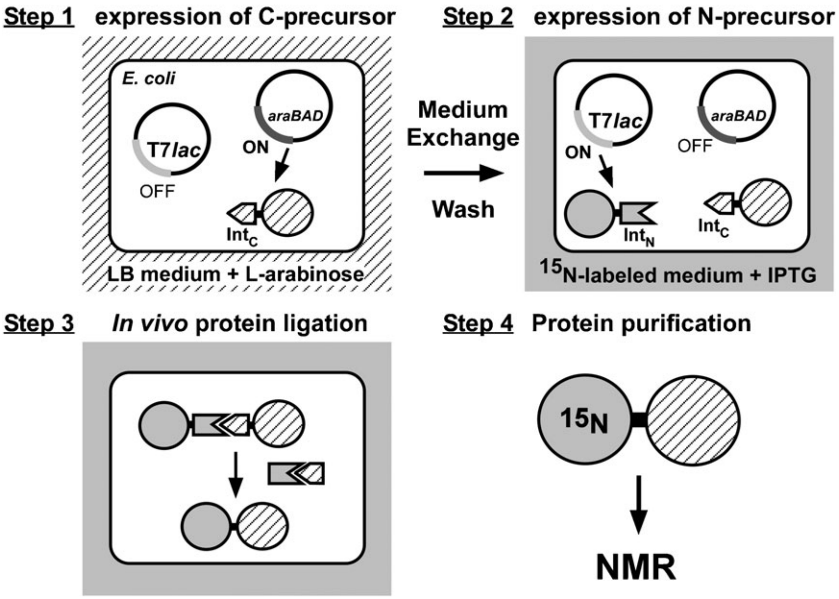
Step 1 expression of C-precursor
E. coli
T7lac
OFF
araBAD
ON
Int$_C$
LB medium + L-arabinose
Step 2 expression of N-precursor
Medium Exchange
Wash
T7lac
ON
araBAD
OFF
Int$_N$
Int$_C$
^{15}N-labeled medium + IPTG
Step 3 In vivo protein ligation
Step 4 Protein purification
15N
NMR

Fig. 5 *In vivo s*egmental isotope labeling using protein *trans*-splicing. The target protein is separated into two segments, each fused to one part of a split intein (Int$_N$ and Int$_C$, respectively). The genes for these two precursor proteins are present on separate plasmids, and expression can be induced with two different small molecules. First, the C-terminal precursor protein is induced with L-arabinose in unlabeled medium (step 1), followed by an exchange of the cells into medium containing ^{15}N. In this isotope-containing medium, the N-terminal precursor protein is induced with IPTG, resulting in protein ligation between the ^{15}N-labeled N-terminal and the unlabeled C-terminal fragment of the target protein (steps 2 and 3). The segmental isotopic labeled full-length target protein is then purified from the cell culture for NMR spectroscopy (step 4). Variations of this approach are possible by expressing only the N-terminal precursor protein in labeled medium, or by including different isotopes during the two induction steps.

protein, the two split inteins must not cross-react with one another in order to avoid undesired products (an AC fusion protein and/or a cyclized B protein).

Central fragment ligation for NMR spectroscopy was first shown to be feasible using orthogonal, artificially split PI-*Pfu*I and PI-*Pfu*II inteins *in vitro* to generate a maltose binding protein, which carried ^{15}N only in an internal segment (residues 101–238 of 370 residues in total).[55] However, since three-fragment ligation requires ligation at two sites, the problem such as lower yield can be significantly magnified. As it is generally known that split inteins result in a higher yield of ligation product *in vivo*, allowing one of the two ligation steps to be carried out *in vivo* provides a means to circumvent this problem. In a seminal report, two non-cross reacting *Npu* DnaE split inteins[9] were used to prepare a multi-domain protein containing the three sequential curacin A acyl carrier protein (ACP) domains (T1, T2, and T3).[40] The protocol first involved the *in vivo* ligation of ^{15}N-labeled T2 to unlabeled T3, followed by *in vitro* ligation of the segmentally labeled T2–T3 to unlabeled T1, thereby producing a modular T1–T2–T3 protein with only the internal T2 domain containing ^{15}N labels. This procedure produced not only simpler NMR spectra than a full-length, uniformly labeled T1–T2–T3 reference protein, but also made it possible to unambiguously assign certain residues to individual T domains.[40] This method was highly significant since the three ACP domains of curacin A are very similar in sequence. In the future, implementation of a central fragment labeling strategy that works entirely *in vivo* is desirable, as this will likely provide the simplest way of preparing such proteins for structural studies.

4. Site-specific fluorescent labeling

Fluorescent probes and proteins have revolutionized cell biology because they allow for visualization of specific molecules inside cells or whole organisms, making it possible to extract information by cellular imaging techniques. But fluorophores are also useful for biophysical analyses of proteins and other biomolecules outside of the cellular context. For example, there is an increasing interest to study the folding of isolated proteins using fluorescence resonance energy transfer at the single-molecule level (smFRET), rather than looking at an ensemble of folding events from a large number of molecules at a given time.[56] The proteins investigated by smFRET so far were of low molecular weight and were easily labeled on native or engineered cysteine residues with fluorescent probes using standard maleimide labeling chemistry.[57–62] Similar studies of larger proteins will likely require other labeling approaches, if native cysteines are inaccessible to labeling and/or introduction of cysteine residues is otherwise problematic.

Split inteins offer a unique opportunity for the incorporation of FRET donor and acceptor molecules into a protein sequence (Fig. 6a). The approach is based on the non-canonical *Ssp* DnaB S1 and *Ssp* GyrB S11 split inteins, which have been shown to efficiently catalyze fluorescent labeling of proteins.[15,16,63,64] The target protein is fused at its N-terminus to the *Ssp* DnaB S1 C-intein (Int1$_C$ in Fig. 6a), and at its C-terminus to the *Ssp* GyrB S11 N-intein (Int2$_N$). Production of the remaining short intein fragments (Int1$_N$ and Int2$_C$, respectively) is accomplished by solid-phase peptide synthesis, allowing for the incorporation of desired fluorophores for

Published on 31 August 2010 on http://pubs.rsc.org | doi:10.1039/C0MB00034E
Downloaded on 01 September 2010

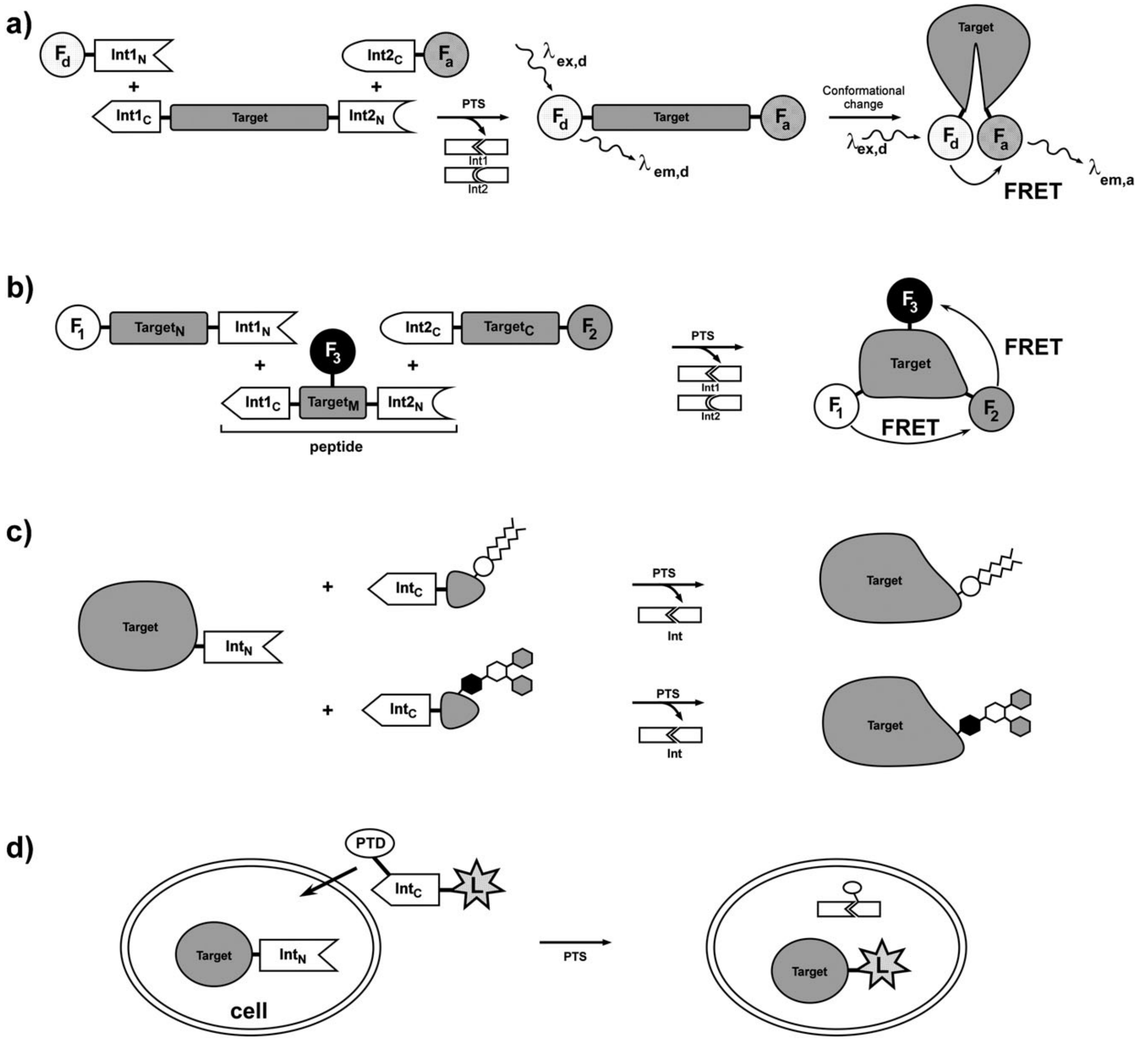

Fig. 6 Potential uses of split inteins in structural biology. (a) Dual-fluorescent labeling of a protein using two split inteins (*e.g.* Int1: *Ssp* DnaB S1,[15,63] Int2: *Ssp* GyrB S11)[16] for fluorescence resonance energy transfer (FRET) studies of protein dynamics. The fluorophores can be attached to the short $Int1_N$ and $Int2_C$ sequences during chemical synthesis of the peptides. F: fluorophore, λ_{ex}: excitation wavelength, λ_{em}: emission wavelength, subscripts 'd' and 'a' refer to donor and acceptor fluorophore, respectively. (b) Central fragment labeling for triple-fluorophore FRET studies. The target protein is divided into three fragments, where the N- and C-terminal parts have been individually labeled at the termini with fluorophores (F_1, F_2) using *e.g.* the scheme outlined in (a), and further contain the large portions of the S11 and S1 split inteins ($Int1_N$ and $Int2_C$, respectively). Incorporating a third fluorophore (F_3) in an internal part of the target protein is achieved by first chemically synthesizing the medial protein sequence ($Target_M$) sandwiched between the short S11 and S1 split intein sequences ($Int1_C$ and $Int2_N$, respectively), and assembly of the full-length, triply-labeled target protein by protein *trans*-splicing. Conformational changes can then be probed either by FRET between fluorophores F_1 and F_2 and/or F_2 and F_3, as indicated. (c) Split intein mediated lipidation (top) and glycosylation (bottom) of a protein's C-terminal tail (CTT) region. The CTT is produced synthetically attached to the Int_C of *e.g.* the *Ssp* GyrB S11 split intein, allowing for the incorporation of any desired lipid molecule or sugar moieties in the CTT. Protein *trans*-splicing then generates the lipidated or glycosylated full-length protein, which can further be analyzed by structural or biophysical methods. (d) In-cell labeling for NMR spectroscopy or fluorescence microscopy. The target protein is produced inside a cell fused to the Int_N part of a split intein. The C-terminal split intein part Int_C is produced chemically, and contains a desired labeling group (L) as well as a protein-transduction domain (PTD) to allow entry into the cell. Upon *trans*-splicing, the target protein acquires the label at the C-terminus in a traceless manner.

FRET. The two PTS reactions will produce dual-labeled target protein, which can then be used for smFRET studies. Split inteins have already been used for FRET studies of protein folding,[22] however, the method reported was a three-step process (not counting in purification steps) and used maleimide chemistry for labeling rather than labeled peptides. The approach using two split inteins could thus be advantageous over the latter technique due to fewer steps and avoidance of chemical labeling.

5. Limitations of split intein technology

Although protein *trans*-splicing mediated by split inteins has proven to be a valuable means of producing segmental isotope-labeled proteins, the techniques outlined above and their opportunities for research have yet to be fully exploited. In the following section, we therefore focus on the intrinsic problems commonly encountered with split intein-based protein ligation and possibly remedy to overcome these shortcomings.[53] Native chemical or expressed protein ligation has also been widely used in preparing segmental isotope-labeled proteins, and the reader is referred to an excellent review[65] on advantages and drawbacks of this particular approach.

The studies using artificially split PI-*Pfu* inteins[49–51,55] required purification of at least one precursor protein under denaturing conditions, and subsequent refolding to remove the denaturant. While this procedure worked for the proteins under investigation (αC, MBP, ATPase β subunit), it is generally desirable to express and purify proteins under native conditions because the success of a refolding experiment cannot be predicted on the basis of the protein primary sequence. Simple dialysis against buffer without denaturant often causes proteins to precipitate. Finding an optimal refolding buffer can be a very time-consuming and tedious undertaking given the sheer infinite number of possible buffer compositions. Another drawback of using the split PI-*Pfu* inteins for protein ligation is the temperature-dependence for efficient ligation. Since the inteins are derived from a thermophilic organism (*Pyrococcus horikoshii*), they catalyze protein ligation most efficiently at elevated temperatures (optimum: 70 °C), but much slower at 37 °C. Hence, in order to be amenable to protein ligation by the PI-*Pfu* split inteins, a target protein must be intrinsically heat-stable, or able to endure long incubations at 37 °C without loss of structural integrity. Therefore, the use of split inteins, which are expressed in a soluble form and which perform efficient protein ligation at temperatures not detrimental to protein stability would be much more favorable over the artificially split PI-*Pfu* inteins. In this respect, it is of note that some artificially split inteins are inherently prone to misfolding when expressed in recombinant form, and thus require subsequent denaturing purification conditions and refolding procedures.[55,66] While this is not true for all artificially split inteins,[67] the naturally occurring DnaE split inteins are superior over genetically engineered split inteins because their naturally split state implies a soluble character. Indeed, both the *Ssp* and *Npu* DnaE split inteins (and derivatives thereof) have been the "gold standard" for protein ligations *in vitro* and *in vivo* because of the advantage of being expressed in a soluble form. Furthermore, both split inteins are naturally found in mesophilic cyanobacteria, and thus efficiently catalyze protein ligation at temperatures of 37 °C or lower.

Another important issue that often hampers the use of split inteins (or inteins in general) for protein ligation experiments is that efficient ligation is dependent on both the splicing junction and the extein sequences. In the following discussion, splicing junction is referred to as the N- and C-terminal intein nucleophiles along with the residues preceding or succeeding them (residues −1 and +2, respectively), whereas extein refers to the entire protein sequence to be ligated. The splicing junction residues are probably important to perfectly align the intein catalytic centers and to provide the chemical environment required for efficient protein splicing without the occurrence of undesired cleavage reactions. The importance of a flexible conformation at the ligation junction is also suggested for PI-*Pfu*I.[50] An Asp residue preceding the N-nucleophile has been reported to most often lead to pronounced levels of premature N-terminal cleavage,[68,69] thereby abolishing protein ligation. Proline at either the −1 or +2 position usually inhibits protein splicing and cleavage completely in some inteins,[38] although this amino acid may occur natively in a few other inteins (*e.g.* Pro-1 in PI-*Pfu*II). The ideal splicing junction is seemingly unique to individual inteins,[38] and the restriction imposed by the splicing junction sequences ultimately influences the choice of where to insert the split intein within a target protein and which intein to use. Table 4 gives a comparison of the split inteins so far reported for mg-quantity preparation of ligated products by PTS and the respective junction residues employed in the ligation. Recently, directed molecular evolution has been used to render inteins less specialized towards their splicing junctions,[70,71] however, no intein is currently available that would splice efficiently in any junction context. To generate such a "super intein", the *Npu* DnaE intein appears to be a good starting point because it was shown to be much more tolerant of different amino acid side chains at the +2 position than the *Ssp* DnaE intein.[38]

Even though this splicing junction dependency exists, one can generally assume that a loop region within a target protein represents a good location for inserting a split intein. Separating the target protein within a loop is less likely to cause detrimental effects on protein folding when it is expressed without the naturally adjacent protein sequence. Loops are also more likely to tolerate mutations that may be necessary to provide the split intein with a favorable splicing function environment. Indeed, in all the studies highlighted above, the split site in the target protein was located in a loop region, and neither mutation nor the addition of extra residues in the loops caused the reconstituted proteins to fold into a structure aberrant from the wild-type protein. Loops may thus be considered a

Table 4 Successfully used splicing junction sequences. The N- and C-nucleophilic residues of the split inteins are in bold. Only additional amino acids inserted between the nucleophiles and the target proteins are shown

Split intein	N-junction	C-junction	Reference
PI-*Pfu*I	GGG/**C**	/**TGL**	49
PI-*Pfu*I	GGG/**C**	/**TGI** /**TGK**	50, 51
PI-*Pfu*II	TNP/**C**	/**CGE**	55
Ssp DnaE	GS/**C**	/**CFNKGT**	54
Npu DnaE	GS/**C**	/**CFNGT**	40
Npu DnaE	TK/**C**	/**CFNG**	76
Npu/*Ssp* DnaE	AEY/**C**	/**CMN**	23

"safe" place for the insertion of split inteins, and simultaneously, the splicing junction dependency may be overcome because mutations and insertions are accommodated more easily. If the protein primary sequence is changed after reconstitution by protein ligation, of course, it has to be ensured that the global structure and function of the protein is not changed, which is generally the case for all structure–function analyses based on mutations.

The more confusing aspect of PTS is the extein dependency in addition to the splicing junction dependency. The efficiency of PTS is modulated by the fused exteins and their order even if an identical splicing junction sequence is used.[9] The mechanism underlying the extein dependency is still unclear and one has to await further investigation. It is currently necessary to assess the feasibility of PTS in each case for a specific intein, preferably, with its ideal junction sequences as a starting point, prior to segmental isotopic labeling.[53]

Central fragment labeling becomes more important and useful for larger proteins, as sites of interest are likely to be distant from both termini. However, the reports published so far using split inteins for protein modification have solely focused on adding labels at either the N- or C-terminus of a target protein.[15,16,20] But central protein modification should be generally possible using the non-canonical S1 and S11 split inteins by chemical synthesis of a medial protein segment with a desired label flanked by the short 6-aa and 11-aa S11 C- and S1 N-intein sequences, and assembly of the full-length target protein by PTS from three fragments (Fig. 6b). Such a scheme could be used *e.g.* to prepare triple-fluorophore labeled proteins for more sophisticated FRET studies of protein conformational changes (Fig. 6b). The current bottleneck of such multiple-fragment ligation by PTS is the ligation efficiencies of individual ligation steps. To obtain sufficient amounts for structural studies, >80–90% of the ligation efficiency at each ligation step is highly desirable, which would still result in a final yield of 60–80% for central labeling. One approach to accomplish high efficiency at all steps is to use only well-characterized highly efficient split inteins.[40]

6. Outlook

Segmental isotopic labeling by multi-fragment ligation exploiting split inteins certainly opens new opportunities for NMR investigation of a domain in intact proteins without dissecting a full-length protein into smaller domains. How else can split inteins help structural biologists apart from their use in segmental isotopic labeling for NMR spectroscopy and producing labeled proteins for single molecule FRET studies? The answer may lie in the ability to use split intein-mediated protein labeling (see Section 4) to incorporate naturally occurring post-translational modifications at desired sites of a protein. Structural investigations of natively modified proteins are often cumbersome because the protein sample can be heterogeneously modified due to the inability to control the degree of post-translational modification inside the native cell. The "manual" addition of such modifications (fatty acids, lipids, sugars) at a specific location in a protein could elegantly be achieved by protein *trans*-splicing. For example, proteins containing lipidated C-terminal tails (CTTs) like Ras play crucial roles in signal transduction processes by passing on incoming signals from the plasma membrane to downstream targets.[72] Lipidation can be achieved by incorporating the desired lipid anchor(s) into the CTT, and fusing this sequence to the short C-intein of *e.g.* the *Ssp* GyrB S11 split intein. Protein *trans*-splicing with the remainder of the target protein (fused to the N-intein sequence) then generates the lipidated full-length protein (Fig. 6c, top), allowing for investigations of the effects of the lipid modification on protein structure and activity. Along these lines, we also see applicability of split inteins to produce recombinant proteins with homogenous glycosylation patterns (Fig. 6c, bottom). In combination with sugar-type specific isotopic labeling,[73] solution NMR spectroscopy of glycosylated proteins—labeled with isotopes both in the protein and the sugars—will likely have an impact to further our understanding of glycosylation structure–function relationships. Most importantly, unlike other chemical approaches, protein *trans*-splicing can generate segmentally or site-specifically labeled native proteins in living cells[19,54] (Fig. 6d), offering high-resolution structural investigation of protein structures *in situ* by fluorescence spectroscopy, cryo-electron tomography, or NMR spectroscopy.[74] Further advances of protein ligation technology by protein *trans*-splicing are likely to provide great opportunities in structural biology, which are limited only by our imagination.

Acknowledgements

The financial support was obtained from the Academy of Finland (1131413). H.I. is Academy Research Fellow of Finland. We thank A. S. Aranko for the unpublished data used in Table 2.

References

1 G. Wider and K. Wüthrich, NMR spectroscopy of large molecules and multimolecular assemblies in solution, *Curr. Opin. Struct. Biol.*, 1999, **9**, 594–601.
2 A. M. Ruschak and L. E. Kay, Methyl groups as probes of supra-molecular structure dynamics and function, *J. Biomol. NMR*, 2010, **46**, 75–87.
3 M. W. Southworth, E. Adam, D. Panne, R. Byer, R. Kautz and F. B. Perler, Control of protein splicing by intein fragment reassembly, *EMBO J.*, 1998, **17**, 918–926.
4 H. Wu, Z. Hu and X. Q. Liu, Protein *trans*-splicing by a split intein encoded in a split DnaE gene of Synechocystis sp. PCC6803, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 9226–9231.
5 F. B. Perler, E. O. Davis, G. E. Dean, F. S. Gimble, W. E. Jack, N. Neff, C. J. Noren, J. Thorner and M. Belfort, Protein splicing elements: inteins and exteins—a definition of terms and recommended nomenclature, *Nucleic Acids Res.*, 1994, **22**, 1125–1127.
6 M. Q. Xu and F. B. Perler, The mechanism of protein splicing and its modulation by mutation, *EMBO J.*, 1996, **15**, 5146–5153.
7 K. Shingledecker, S. Q. Jiang and H. Paulus, Molecular dissection of the Mycobacterium tuberculosis RecA intein: design of a minimal intein and of a *trans*-splicing system involving two intein fragments, *Gene*, 1998, **207**, 187–195.
8 K. V. Mills, B. M. Lew, S. Jiang and H. Paulus, Protein splicing in trans by purified N- and C-terminal fragments of the *Mycobacterium tuberculosis* RecA intein, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 3543–3548.
9 A. S. Aranko, S. Züger, E. Buchinger and H. Iwaï, *In vivo* and *in vitro* protein ligation by naturally occurring and engineered split DnaE inteins, *PLoS One*, 2009, **4**, e5185.

10 J. S. Oeemig, A. S. Aranko, J. Djupsjöbacka, K. Heinämäki and H. Iwaï, Solution structure of DnaE intein from *Nostoc punctiforme*: structural basis for the design of a new split intein suitable for site-specific chemical modification, *FEBS Lett.*, 2009, **583**, 1451–1456.
11 Y. Ding, M. Q. Xu, I. Ghosh, X. Chen, S. Ferrandon, G. Lesage and Z. Rao, Crystal structure of a mini-intein reveals a conserved catalytic module involved in side chain cyclization of asparagine during protein splicing, *J. Biol. Chem.*, 2003, **278**, 39133–39142.
12 P. Sun, S. Ye, S. Ferrandon, T. C. Evans, M. Q. Xu and Z. Rao, Crystal structures of an intein from the split dnaE gene of Synechocystis sp. PCC6803 reveal the catalytic model without the penultimate histidine and the mechanism of zinc ion inhibition of protein splicing, *J. Mol. Biol.*, 2005, **353**, 1093–1105.
13 W. Sun, J. Yang and X. Q. Liu, Synthetic two-piece and three-piece split inteins for protein *trans*-splicing, *J. Biol. Chem.*, 2004, **279**, 35281–35286.
14 J. H. Appleby, K. Zhou, G. Volkmann and X. Q. Liu, Novel Split Intein for *trans*-Splicing Synthetic Peptide onto C Terminus of Protein, *J. Biol. Chem.*, 2009, **284**, 6194–6199.
15 C. Ludwig, M. Pfeiff, U. Linne and H. D. Mootz, Ligation of a synthetic peptide to the N terminus of a recombinant protein using semisynthetic protein *trans*-splicing, *Angew. Chem., Int. Ed.*, 2006, **45**, 5218–5221.
16 G. Volkmann and X. Q. Liu, Protein C-terminal labeling and biotinylation using synthetic peptide and split-intein, *PLoS One*, 2009, **4**, e8381.
17 T. C. Evans, Jr., D. Martin, R. Kolly, D. Panne, L. Sun, I. Ghosh, L. Chen, J. Benner, X. Q. Liu and M. Q. Xu, Protein *trans*-splicing and cyclization by a naturally split intein from the dnaE gene of Synechocystis species PCC6803, *J. Biol. Chem.*, 2000, **275**, 9091–9094.
18 Y. Kwon, M. A. Coleman and J. A. Camarero, Selective immobilization of proteins onto solid supports through split-intein-mediated protein *trans*-splicing, *Angew. Chem., Int. Ed.*, 2006, **45**, 1726–1729.
19 I. Giriat and T. W. Muir, Protein semi-synthesis in living cells, *J. Am. Chem. Soc.*, 2003, **125**, 7180–7181.
20 T. Kurpiers and H. D. Mootz, Regioselective cysteine bioconjugation by appending a labeled cysteine tag to a protein by using protein splicing in trans, *Angew. Chem., Int. Ed.*, 2007, **46**, 5234–5237.
21 T. Kurpiers and H. D. Mootz, Site-specific chemical modification of proteins with a prelabelled cysteine tag using the artificially split Mxe GyrA intein, *ChemBioChem*, 2008, **9**, 2317–2325.
22 J. Y. Yang and W. Y. Yang, Site-specific two-color protein labeling for FRET studies using split inteins, *J. Am. Chem. Soc.*, 2009, **131**, 11644–11645.
23 H. Al-Ali, T. J. Ragan, X. Gao and T. K. Harris, Reconstitution of modular PDK1 functions on *trans*-splicing of the regulatory PH and catalytic kinase domains, *Bioconjugate Chem.*, 2007, **18**, 1294–1302.
24 J. Shi and T. W. Muir, Development of a tandem protein *trans*-splicing system based on native and engineered split inteins, *J. Am. Chem. Soc.*, 2005, **127**, 6198–6206.
25 T. Ozawa and Y. Umezawa, Detection of protein-protein interactions *in vivo* based on protein splicing, *Curr. Opin. Chem. Biol.*, 2001, **5**, 578–583.
26 T. Ozawa and Y. Umezawa, Identification of proteins targeted into the endoplasmic reticulum by cDNA library screening, *Methods Mol. Biol.*, 2007, **390**, 269–280.
27 S. B. Kim, T. Ozawa, S. Watanabe and Y. Umezawa, High-throughput sensing and noninvasive imaging of protein nuclear transport by using reconstitution of split *Renilla luciferase*, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 11542–11547.
28 H. G. Chin, G. D. Kim, I. Marin, F. Mersha, T. C. Evans, Jr., L. Chen, M. Q. Xu and S. Pradhan, Protein *trans*-splicing in transgenic plant chloroplast: reconstruction of herbicide resistance from split genes, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 4510–4515.
29 T. C. Evans, Jr., M. Q. Xu and S. Pradhan, Protein splicing elements and plants: from transgene containment to protein purification, *Annu. Rev. Plant Biol.*, 2005, **56**, 375–392.
30 J. Li, W. Sun, B. Wang, X. Xiao and X. Q. Liu, Protein *trans*-splicing as a means for viral vector-mediated *in vivo* gene therapy, *Hum. Gene Ther.*, 2008, **19**, 958–964.
31 S. Brenzel, M. Cebi, P. Reiss, U. Koert and H. D. Mootz, Expanding the scope of protein *trans*-splicing to fragment ligation of an integral membrane protein: towards modulation of porin-based ion channels by chemical modification, *ChemBioChem*, 2009, **10**, 983–986.
32 C. P. Scott, E. Abel-Santos, M. Wall, D. C. Wahnon and S. J. Benkovic, Production of cyclic peptides and proteins *in vivo*, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 13638–13643.
33 H. Iwai, A. Lingel and A. Plückthun, Cyclic green fluorescent protein produced *in vivo* using an artificially split PI-PfuI intein from *Pyrococcus furiosus*, *J. Biol. Chem.*, 2001, **276**, 16548–16554.
34 H. Iwai and A. Plückthun, Circular beta-lactamase: stability enhancement by cyclizing the backbone, *FEBS Lett.*, 1999, **459**, 166–172.
35 N. K. Williams, P. Prosselkov, E. Liepinsh, I. Line, A. Sharipo, D. R. Littler, P. M. Curmi, G. Otting and N. E. Dixon, *In vivo* protein cyclization promoted by a circularly permuted Synechocystis sp. PCC6803 DnaB mini-intein, *J. Biol. Chem.*, 2002, **277**, 7790–7798.
36 G. Volkmann, P. W. Murphy, E. E. Rowland, J. E. Cronan, Jr., X. Q. Liu, C. Blouin and D. M. Byers, Intein-mediated cyclization of bacterial acyl carrier protein stabilizes its folded conformation but does not abolish function, *J. Biol. Chem.*, 2010, **285**, 8605–8614.
37 P. E. Dawson, T. W. Muir, I. Clark-Lewis and S. B. Kent, Synthesis of proteins by native chemical ligation, *Science*, 1994, **266**, 776–779.
38 H. Iwai, S. Züger, J. Jin and P. H. Tam, Highly efficient protein *trans*-splicing by a naturally split DnaE intein from *Nostoc punctiforme*, *FEBS Lett.*, 2006, **580**, 1853–1858.
39 B. Dassa, G. Amitai, J. Caspi, O. Schueler-Furman and S. Pietrokovski, Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations, *Biochemistry*, 2007, **46**, 322–330.
40 A. E. Busche, A. S. Aranko, M. Talebzadeh-Farooji, F. Bernhard, V. Dötsch and H. Iwaï, 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.*, 2009, **48**, 6128–6131.
41 R. A. Venters, R. Thompson and J. Cavanagh, Current approaches for the study of large proteins by NMR, *J. Mol. Struct.*, 2002, **602**, 275–292.
42 R. Xu, B. Ayers, D. Cowburn and T. W. Muir, Chemical ligation of folded recombinant proteins: segmental isotopic labeling of domains for NMR studies, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 388–393.
43 J. A. Camarero, A. Shekhtman, E. A. Campbell, M. Chlenov, T. M. Gruber, D. A. Bryant, S. A. Darst, D. Cowburn and T. W. Muir, Autoregulation of a bacterial sigma factor explored by using segmental isotopic labeling and NMR, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 8536–8541.
44 K. J. Walters, P. J. Lech, A. M. Goh, Q. Wang and P. M. Howley, DNA-repair protein hHR23a alters its protein structure upon binding proteasomal subunit S5a, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 12694–12699.
45 F. Vitali, A. Henning, F. C. Oberstrass, Y. Hargous, S. D. Auweter, M. Erat and F. H. Allain, Structure of the two most C-terminal RNA recognition motifs of PTB using segmental isotope labeling, *EMBO J.*, 2006, **25**, 150–162.
46 L. Skrisovska and F. H. Allain, Improved segmental isotope labeling methods for the NMR study of multidomain or large proteins: application to the RRMs of Npl3p and hnRNP L, *J. Mol. Biol.*, 2008, **375**, 151–164.
47 W. Zhao, Y. Zhang, C. Cui, Q. Li and J. Wang, An efficient on-column expressed protein ligation strategy: application to segmental triple labeling of human apolipoprotein E3, *Protein Sci.*, 2008, **17**, 736–747.
48 P. S. Hauser, V. Raussens, T. Yamamoto, G. E. Abdullahi, P. M. Weers, B. D. Sykes and R. O. Ryan, Semisynthesis and segmental isotope labeling of the apoE3 N-terminal domain using expressed protein ligation, *J. Lipid Res.*, 2009, **50**, 1548–1555.
49 T. Yamazaki, T. Otomo, N. Oda, Y. Kyogoku, K. Uegaki, N. Ito, Y. Ishino and H. Nakamura, Segmental isotope labeling for protein NMR using peptide splicing, *J. Am. Chem. Soc.*, 1998, **120**, 5591–5592.

50 T. Otomo, K. Teruya, K. Uegaki, T. Yamazaki and Y. Kyogoku, Improved segmental isotope labeling of proteins and application to a larger protein, *J. Biomol. NMR*, 1999, **14**, 105–114.

51 H. Yagi, T. Tsujimoto, T. Yamazaki, M. Yoshida and H. Akutsu, Conformational change of H+-ATPase beta monomer revealed on segmental isotope labeling NMR spectroscopy, *J. Am. Chem. Soc.*, 2004, **126**, 16632–16638.

52 M. Muona, A. S. Aranko and H. Iwaï, Segmental isotopic labelling of a multidomain protein by protein ligation by protein *trans*-splicing, *ChemBioChem*, 2008, **9**, 2958–2961.

53 M. Muona, A. S. Aranko, V. Raulinaitis and H. Iwaï, Segmental isotopic labeling of multi-domain and fusion proteins by protein *trans*-splicing *in vivo* and *in vitro*, *Nat. Protoc.*, 2010, **5**, 574–587.

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

55 T. Otomo, N. Ito, Y. Kyogoku and T. Yamazaki, NMR observation of selected segments in a larger protein: central-segment isotope labeling through intein-mediated ligation, *Biochemistry*, 1999, **38**, 16040–16044.

56 B. Schuler and W. A. Eaton, Protein folding studied by single-molecule FRET, *Curr. Opin. Struct. Biol.*, 2008, **18**, 16–26.

57 E. V. Kuzmenkina, C. D. Heyes and G. U. Nienhaus, Single-molecule FRET study of denaturant induced unfolding of RNase H, *J. Mol. Biol.*, 2006, **357**, 313–324.

58 T. Tezuka-Kawakami, C. Gell, D. J. Brockwell, S. E. Radford and D. A. Smith, Urea-induced unfolding of the immunity protein Im9 monitored by spFRET, *Biophys. J.*, 2006, **91**, L42–L44.

59 A. A. Deniz, T. A. Laurence, G. S. Beligere, M. Dahan, A. B. Martin, D. S. Chemla, P. E. Dawson, P. G. Schultz and S. Weiss, Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 5179–5184.

60 K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich and W. A. Eaton, Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 1528–1533.

61 T. A. Laurence, X. Kong, M. Jager and S. Weiss, Probing structural heterogeneities and fluctuations of nucleic acids and denatured proteins, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 17348–17353.

62 F. Huang, S. Sato, T. D. Sharpe, L. Ying and A. R. Fersht, Distinguishing between cooperative and unimodal downhill protein folding, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 123–127.

63 C. Ludwig, D. Schwarzer and H. D. Mootz, 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.*, 2008, **283**, 25264–25272.

64 T. Ando, S. Tsukiji, T. Tanaka and T. Nagamune, Construction of a small-molecule-integrated semisynthetic split intein for *in vivo* protein ligation, *Chem. Commun.*, 2007, 4995–4997.

65 D. Cowburn and T. W. Muir, Segmental isotopic labeling using expressed protein ligation, *Methods Enzymol.*, 2001, **339**, 41–54.

66 B. M. Lew, K. V. Mills and H. Paulus, Protein splicing *in vitro* with a semisynthetic two-component minimal intein, *J. Biol. Chem.*, 1998, **273**, 15887–15890.

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

68 M. W. Southworth, K. Amaya, T. C. Evans, M. Q. Xu and F. B. Perler, Purification of proteins fused to either the amino or carboxy terminus of the *Mycobacterium xenopi* gyrase A intein, *BioTechniques*, 1999, **27**, 110–114116, 118–120.

69 G. Amitai, B. P. Callahan, M. J. Stanger, G. Belfort and M. Belfort, Modulation of intein activity by its neighboring extein substrates, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 11005–11010.

70 S. W. Lockless and T. W. Muir, Traceless protein splicing utilizing evolved split inteins, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 10999–11004.

71 K. Hiraga, I. Soga, J. T. Dansereau, B. Pereira, V. Derbyshire, Z. Du, C. Wang, P. Van Roey, G. Belfort and M. Belfort, Selection and structure of hyperactive inteins: peripheral changes relayed to the catalytic center, *J. Mol. Biol.*, 2009, **393**, 1106–1117.

72 K. Wennerberg, K. L. Rossman and C. J. Der, The Ras superfamily at a glance, *J. Cell Sci.*, 2005, **118**, 843–846.

73 L. Skrisovska, M. Schubert and F. H. Allain, Recent advances in segmental isotope labeling of proteins: NMR applications to large proteins and glycoproteins, *J. Biomol. NMR*, 2009, **46**, 51–65.

74 D. Sakakibara, A. Sasaki, T. Ikeya, J. Hamatsu, T. Hanashima, M. Mishima, M. Yoshimasu, N. Hayashi, T. Mikawa, M. Walchli, B. O. Smith, M. Shirakawa, P. Güntert and Y. Ito, Protein structure determination in living cells by in-cell NMR spectroscopy, *Nature*, 2009, **458**, 102–105.

75 E. C. Johnson and S. B. Kent, Insights into the mechanism and catalysis of the native chemical ligation reaction, *J. Am. Chem. Soc.*, 2006, **128**, 6640–6646.

76 E. Buchinger, F. L. Aachmann, A. S. Aranko, S. Valla, G. Skjak-Braek, H. Iwaï and R. Wimmer, Use of protein *trans*-splicing to produce active and segmentally (2)H, (15)N labelled mannuronan C5-epimerase AlgE4, *Protein Sci.*, 2010, **19**, 1534–1543.