Published in final edited form as:

Curr Protein Pept Sci. 2019 ; 20(5): 408–424. doi:10.2174/1389203720666190208110416.

Biotechnological applications of protein splicing

Corina Sarmiento¹, Julio A. Camarero^{1,2,3}

¹Department of Pharmacology and Pharmaceutical Sciences, University of Southern California, Los Angeles, CA9033 USA,

²Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA9033 USA,

³Department of Chemistry, University of Southern California, Los Angeles, California 90089-9121, USA,

Abstract

Protein splicing domains, also called inteins, have become a powerful biotechnological tool for applications involving molecular biology and protein engineering. Early applications of inteins focused on self-cleaving affinity tags, generation of recombinant polypeptide α -thioesters for the production of semisynthetic proteins and backbone cyclized polypeptides. The discovery of naturally occurring split-inteins has allowed the development of novel approaches for the selective modification of proteins both *in vitro* and *in vivo*. This review gives a general introduction to protein splicing with a focus on their role in expanding the applications of intein-based technologies in protein engineering and chemical biology.

Graphical Abstract

Keywords

Inteins; split-inteins; protein modification; protein trans-splicing; protein splicing; post-translational modification; conditional protein splicing; backbone cyclized proteins

1. Introduction

Inteins are protein domains found in nature with the capability to carry out the process of protein splicing. Protein splicing is a post-translation biochemical modification which results in the cleavage and formation of peptide bonds between precursor polypeptide segments flanking the intein. Protein splicing excises the internal region of the precursor protein, which is then followed by the ligation of the N-extein and C-extein fragments, resulting in two polypeptides, the excised intein and the new polypeptide produced by joining the C- and N-extended [1] (Fig. 1A). Although in many inteins this biochemical reaction is spontaneous not requiring any external factor or energy source to work although, the activity of some naturally-occurring and engineered inteins can be triggered by changes in temperature, pH, the presence of other biomolecules and oxidative stress [2].

The first protein splicing domain or intein was discovered in the vacuolar proton-translocating ATPase gene of *Saccharomyces cerevisiae* (*Sce* VMA1) in 1990 [3]. Since then, there have been more than 600 presumed intein genes identified in the genome of many unicellular organisms, such as eukaryote, bacteria, archaea and viruses [4]. In nature inteins are in-frame intervening sequences that modify a host gene and its protein product [1]. Many inteins have bifunctional activities, with separate domains responsible for homing endonuclease activities and protein splicing. Over 75% of the inteins have been found to contain either a DOD or H-N-H class homing endonuclease domains [5]. Intein homing endonucleases can facilitate the mobilization of its own coding sequence to alleles that do not contain such sequences (see references [5c, 6] for recent reviews on intein homing endonucleases). Inteins are usually inserted into proteins involved in DNA replication, transcription and housekeeping genes [7]. Protein splicing domains are usually inserted within conserved protein sequences in their host [8]. The location of these domains within regions essential for protein function initially led people to believe intein excision was required for host protein function. Interestingly, no intein has clearly been shown to be essential in the regulation of its host protein or its host organism [9]. The redox-controlled splicing shown by some inteins *in vitro* has led to speculate that these inteins may inhibit the activity of some endogenous host protein's assembly and function under hyperoxic conditions, however this has not been validated yet in a native host organism [10].

Inteins can be divided into three subgroups: full-length, mini and split inteins. Split inteins make up less than 5% of all the intein genes identified so far [4]. Split inteins are unique in that they are transcribed and translated on two separate genes. Once translated, the intein fragments bind to each other allowing them to assemble intermolecularly into a fully functional protein splicing unit, which catalyzes protein splicing in *trans*. In contrast to the typical *cis* intein, which is composed of an N-extein and a C-extein connected by the intein domain (Fig. 1B), split inteins are split into two fragments called N-intein (I_N) and C-intein (I_C). The I_N and I_C fragments are fused to the C- and N-termini of the N-extein and C-extein polypeptides, respectively (Fig. 1B). Split inteins commonly have an I_N sequence containing around 120 residues in length, while I_C fragments are smaller containing around 40 amino acids [11]. The first naturally occurring split intein identified was found in the alpha subunit of the DNA polymerase III (DnaE) protein from cyanobacterium *Synechocystis sp* (*Ssp* DnaE) [12]. Similar split inteins have also been found in other cyanobacteria [13]. Of

particular relevance is the DnaE split intein found in the cyanobacterium *Nostoc punctiforme* (*Npu* DnaE), which is significantly more efficient in splicing than the *Ssp* DnaE split intein [14]. The DnaE split inteins have been extensively used for the modification of proteins [15]. Artificial split inteins can be also created from naturally occurring *cis* inteins. This has allowed the engineering of split inteins with shorter I_N and I_C fragments to facilitate the production of chemically modified polypeptides at the C- and N-termini [15]. These artificially split inteins are usually less efficient than the naturally occurring split inteins [16].

The different applications of inteins have been extensively reviewed in the recent literature [7a, 15, 17]. Here, we focus on describing in detail the most recent intein biotechnological applications in molecular biology and protein engineering, including protein expression and modification, post-translational processing and labeling, protein regulation by conditional protein splicing, biosensors, and the expression of trans-genes.

2. Mechanism of protein splicing

The canonical mechanism in which inteins excise themselves from their precursor protein is through a sequence of acyl-transfer reactions, resulting in the cleavage of two peptide bonds between the intein and extein regions and the formation of a peptide bond between the N-extein and C-extein [18]. The different steps during the protein splicing reaction are facilitated by the presence of specific shared conserved amino acid residues. Although inteins share a low sequence homology (less than 40%) they do share conserved residues, primarily at their N- and C-termini. Most inteins contain either a Cys or Ser residue at its N-terminus and a His-Asn at their C-terminus. There are also several conserved residues within the intein sequence such as the Thr and His residues found in the TXXH motif [1, 19]. The first amino acid of the C-extein is also highly conserved and must be a residue containing a hydroxyl or thiol group in position β , such as Cys, Thr or Ser (Fig. 2) [4].

The accepted canonical mechanism for protein splicing involves four steps (Fig. 3). First, the thiol/alcohol group of the first residue of the intein, a Cys/Ser residue, attacks the peptide bond at the junction between the N-extein and the intein resulting in the formation of linear (thio)ester intermediate (Fig. 3). This N to O/S acyl transfer is facilitated by a conserved Thr and His residues located in the TXXH motif of the intein. This (thio)ester intermediate then undergoes a trans(thio)esterification reaction by nucleophilic attack of the first C-extein residue, which is invariably a Cys, Ser, or Thr. The resulting branched (thio)ester intermediate undergoes a unique transformation involving the cyclization of the highly conserved C-terminal Asn residue of the intein resulting in the formation of C-terminal succinimide. This process is facilitated by the penultimate His residue and excises the intein from the reactive complex leaving behind the exteins attached through a non-peptidic linkage, which rapidly rearranges into a stable peptide bond in an intein-independent fashion [20].

Intriguingly, some inteins have slightly divergent sequences in the canonical splicing motifs which require alternate mechanisms of splicing. For example, some inteins lack the Cys/Ser residue at the first position of the intein sequence, having instead an Ala or Pro as their N-

terminal residue. Consequently, they cannot produce the first linear (thio)ester intermediate and they must directly form the typical branched intermediate upon N-terminal activation [21], or proceed through a different branched thioester intermediate using a cysteine residue located at the Block F before the canonical branched structure is formed. In addition, a few inteins have also been found to have either Gln or Asp, rather than Asn at the last residue of the intein. Theoretically, these inteins could proceed through a similar mechanism involving the cleavage and formation of a C-terminal glutarimide or succinic anhydride, respectively, rather than a succinimide [22].

3. Biotechnological applications of protein splicing

Inteins provide a powerful and versatile protein engineering tool, which can be utilized *in vitro* to chemically modify polypeptides. Inteins have been used for protein engineering, protein labeling, protein cyclization and protein purification. The first major use of inteins in protein purification was the development of self-cleaving affinity tags to produce an untagged protein [23]. Modified inteins were also initially used for the production of recombinant C-terminal polypeptide α -thioesters, which are key components for the semi-synthesis of chemically modified proteins using expressed protein ligation (EPL) [24]. EPL is used to chemically ligate synthetic peptides to a recombinant C-terminal protein α -thioester using a standard native chemical ligation (NCL) [25]. Protein trans-splicing (PTS) mediated by naturally-occurring or artificially designed split-inteins has been also widely used for the modification of proteins both *in vitro* and in cell [17a, 26]. Inteins have been used to introduce site specific modifications into proteins, such as phosphorylation, biotinylation, ubiquitination, glycosylation, and segmental isotopic labeling [27]. In addition, PTS has also been used *in vivo* to add synthetic probes, such as noncoded fluorescent probes [28]. Inteins can also be modified to make intein-based biosensors [29]. The discovery and implementation of highly efficient split inteins in the last few years has led to a drastic increase in use of intein-based applications, making split inteins a novel chemical and biological tool. In the next sections, we will review in detail the most recent biotechnological applications of inteins, which can be considered a gift from nature to the chemical biologist.

3.1 Protein purification

Protein purification is a key component for many biotechnological and research applications. The need for novel methods able to yield highly purified recombinant at large scale in a relatively non-expensive fashion has led to the development of different self-processing intein-based affinity-tag purification techniques [17a, 26]. Affinity tags were originally designed to be removed by specific endoproteases. The use of endoproteases is a costly factor in the process of protein purification and more importantly it may interfere with the biological activity of the purified protein [30]. This issue has been largely overcome by the use of affinity tags in combination with self-cleaving modified intein domains [17a, 26, 31]. In order to design a self-cleaving affinity tag, a modified intein containing an appropriate affinity tag must be fused to the protein of interest (POI). Once the tagged protein is affinity-purified, the tag can be removed by inducing the intein to cleave (Fig. 4). Interruption of certain steps during the splicing reaction by point mutation can convert the splicing reaction

into a cleavage reaction at either splice junction [23b, 32]. For example, mutation of the first and last residue in the intein leads to C- and N-terminal cleavage, respectively.

The first self-cleaving affinity tag termed IMPACT system was developed in 1997 by New England Biolabs [27c, 33]. The IMPACT system utilizes the engineered *Sce* VMA1 intein, which requires the addition of a thiol-containing compound to activate the N-terminal cleavage of the intein. N-terminal cleavage is favored by mutation of the C-terminal Asn residue to Ala in the intein. This mutation inhibits progression of the splicing reaction after the initial N->S acyl shift, therefore trapping the linear thioester intermediate and favoring N-terminal cleavage upon addition of a soluble thiol-containing compound through a trans-thioesterification reaction. In this system the POI is fused to the N-terminus of the engineered intein. To facilitate affinity purification of the POI-intein fusion a chitin binding is added at the C-terminus of the intein. Hence, once the protein is purified using a chitin-sepharose column, cleavage of the POI can be easily induced by thiols like 1,4-dithiothreitol (DTT) or β -mercaptoethanol in buffers at physiological pH. It should be noted, however, that if the cleaved POI is required to perform α -thioester-based ligations, a thiol not containing any heteroatom in position β should be used instead, e.g. sodium 2-mercaptoethanesulfonate (MESNA) [34]. The development of the IMPACT-CN system permitted N- and C-terminal cleavage of the intein [23c], allowing fusion of the self-cleaving tagged intein to either the N or C-termini of the POI.

The discovery of more rapid C-terminal cleaving inteins led to the development of new intein-tag purification systems. These inteins showed suppressed cleavage activity at pH 8.5, allowing them to be purified on affinity columns at this pH and then cleaved at pH 6.5, releasing the cleaved POI from the column. These inteins also showed strong regulation by temperature, with the highest cleavage activity observed at 37° C [2]. These new findings allowed New England Biolabs to develop a two-intein purification system based on the pTWIN-vector [35]. This expression system is designed for either protein purification or for the isolation of proteins with an N-terminal cysteine and/or a C-terminal α -thioester. The expression vector contains a polylinker region that is designed for the in-frame fusion of a target gene between the modified *Ssp* DnaB [2a] and *Mxe* GyrA inteins [23a, 33a]. The C-terminal cleavage of the *Ssp* DnaB intein can be regulated through temperature and/or pH [2a], while the N-terminal cleavage of the *Mxe* GyrA intein can be induced by addition of a thiol compound [23a].

The development of these intein-based protein expression systems has led to a variety of new tag systems and configurations for protein purification, which initially included the use of chitin- [23c], maltose-binding proteins [36], and small ubiquitin-like modifier (SUMO) and ubiquitin tags [37]. More interesting is the combination of inteins and non-chromatographic purification tags using elastin-like polypeptides (ELPs) and a phasin/polyhydroxybutyrate (PHB) system for the cost-efficient production and purification of recombinant proteins [38]. ELP-based tags can self-associate forming insoluble aggregates under high salt conditions, which can be then separated by centrifugation [38b]. This process can be repeated several times to purify the ELP-intein-fusion containing the POI. Once purified, the intein-ELP tag can be cleaved yielding the POI. Aggregation of this expression system can be controlled through concentration, temperature, length of the ELP repeats

and/or the type of salt [39]. The low cost, convenience and high output of this ELP-intein-based system makes it a powerful tool for protein purification. The PHB system uses a multiple phasin tag produced by *Cupriavidus necator* bacteria, which binds to polyhydroxybutyrate granules in *E. coli* [38a, 40]. The intein fused phasin-tagged protein of interest can be expressed in *E. coli* producing polyhydroxybutyrate granules. The binding of the tag to polyhydroxybutyrate granules allows the separation of the protein of interest from the crude extract after centrifugation. Intein cleavage is induced, severing the protein of interest from the phasin-tag and from the PHB granules, resulting in a highly purified protein [38a, 41].

The ability to minimize cleavage of the intein fusion protein during protein expression and lysis of the cell, but at the same time be able to cleave the intein rapidly once the fusion precursor is purified is a key requirement for all intein-based purification systems. The use of split-inteins should allow this by controlling intein function through reassembly of the *trans*-cleaving split-inteins [17a]. Split-inteins have their key catalytic residues sites located on two different polypeptide fragments, I_C and I_N, and therefore their activity can be only initiated when both fragments assemble intermolecularly producing an active intein in *trans*. When split inteins are used for purification purposes, the POI is fused to either the I_C or I_N intein fragments containing an appropriate purification tag. Once this fusion protein is purified, addition of the complementary intein fragment triggers the cleavage and release of the POI.

The artificially split *Ssp* DnaB mini intein was the first to be used in a split intein-based purification system [16e]. In order to suppress premature cleavage during expression the *Ssp* DnaB intein was engineered by deleting the first 11-amino acids at its N-terminus. This deletion inactivates its splicing activity allowing the purification of a C-terminally fused target protein. The cleavage is induced by adding the remaining 11-residue peptide releasing the POI from the immobilized I_C-tag fragment during purification. The excess of the I_N peptide is usually removed by dialysis. Although this system is efficient and reduces premature cleavage of the POI, the need to use a synthetic peptide (I_N fragment) and an extra dialysis process makes this approach less attractive.

Another split intein based purification system also used *Ssp* DnaB intein with a different split junction [42]. In this system, the tagged I_C or I_N fragments were first bound to the chromatographic column, and then the corresponding POI-I_N or I_C-POI were then added to the column leaving the inteins fragments to bind each other at pH 5. Increasing the pH to 6.5 activates the intein cleavage, releasing the POI while the intein fragments remain bound to the column. Although this system is efficient, it usually requires long cleavage reaction times.

The naturally-occurring *Npu* DnaE split-intein, which exhibits extremely rapid *trans*-splicing [28, 43], has been also used for the design of an efficient split intein purification system [44]. This intein was modified by introducing the D118G mutation at the I_N fragment, which uncouples the N- and C-terminal cleavage activity, suppressing N-terminal cleavage while still exhibiting a rapid C-terminal cleave [44].

An artificially split intein version of the *Mtu* mini-RecA intein has been also used for protein purification purposes [45]. In this system both intein fragments, I_C and I_N, are fused to the C-terminus of an ELP-based tag, while the POI is fused to the C-terminus of the I_C intein fragment. Individually each intein segment is incapable of triggering cleavage, but when combined the assembled intein cleaves the target protein rapidly. This results in a very tight control of the reaction cleaving the tag, which combined with the simplicity of the ELP purification method, yields a very powerful system for the rapid purification of proteins.

3.2 Site-specific labeling of proteins

Protein function is frequently regulated by post-translation modifications (PTMs). Hence, production of site-specifically modified proteins is of extreme importance in order to decipher the structure/function of proteins. Intein-mediated EPL and PTS have been widely used for the site-specific modification of proteins, including the introduction of phosphoryl, glycosyl, biotinyl, ubiquitinyl groups, as well as the production of segmental isotopic labeled proteins, among others (see Refs. [15, 17a, 26] for recent reviews).

Since its introduction in the late 90s, EPL has facilitated the structural study of proteins containing specific PTMs. The use of standard protein expression methods to produce protein samples containing chemically well-defined PTMs is extremely challenging. Early reports, described the production of homogenous samples of the TFG- β signaling proteins, Smad2 and Smad3, using EPL [46].

These mammalian transcription factors can oligomerize upon phosphorylation to be translocated into the nucleus in order to control transcription. The preparation of highly homogeneous Smad proteins that are phosphorylated at specific Ser and Thr residues was key to obtain several high-resolution crystal structures of these proteins [46–47].

Prenylation is another type of PTM found in eukaryotic proteins to control their biological function. For example, prenylation is found in Rab GTPases, which are anchored to the membrane through the lipids introduced at its C-termini, playing a critical role in membrane trafficking. The site-specific introduction of lipids into proteins is extremely challenging due to the low stability and solubility of lipidated proteins. EPL has been used to prepare mono- and bi-prenylated versions of the small Rab/Ypt protein family member Ypt1 [48]. The crystal structures of the lipidated versions of these proteins bound to the Rab GDP dissociation inhibitor (RabGDI) were obtained [49]. These results provided valuable structural information on how RabGDI is able to extract lipidated-Rabs from membranes to facilitate their translocation to acceptor membranes [49].

EPL has also made important contributions to understand the biological effects of PTMs on the biological activities of histones on gene regulation [50]. Histones are heavily post-translationally modified at their flexible N-terminal tails. The regio-selective introduction of these PTMS is a key element for the regulation of both the function and structure of chromatin in DNA replication, transcription, repair and condensation. The production of homogeneously modified histones has been accomplished by several groups using protein ligation techniques [51].

More interesting is the application of inteins for the site-specific labeling of proteins in living cells. PTS can be used for this task quite efficiently by using split inteins formed by either short I_N or I_C fragments. This facilitates their chemical synthesis as well as the introduction of appropriate chemical probes. The use of PTS for this task presents several advantages over other chemoselective ligation techniques. For example, the PTS-mediated labeling process relies on a specific molecular recognition event between the corresponding I_N and I_C fragments. This should allow simultaneous multicolor labeling of proteins in live cells by making use of orthogonal split inteins, *i.e.* split inteins that do not show cross-reactivity with each other [52]. It is worth noting that using PTS for the site-specific labeling of proteins with fluorescent probes inside live cells requires that the labeling process be linked to the simultaneous activation of fluorescence. This is critical to minimize the background signal from the unreacted intein labeled fragment. The use of a Förster resonance energy transfer (FRET) quenched split intein was used to accomplish this task [28]. In this work, the fluorogenic probe was site-specifically introduced at the C terminus of the *Npu* DnaE C-extein. To quench the dye fluorescence a specific quencher was also introduced on the I_C intein to allow quenching the dye fluorescence in the unreacted I_C fragment (Fig. 5A). Once the PTS reaction occurs the fluorescent dye is ligated to the POI, while simultaneously dissociating the fluorophore from its quencher and activating its fluorescence. This elegant approach has been successfully employed for the site-specific in-cell labeling of the DNA-binding domain (DBD) of the transcription factor YY1 in several cell lines [28]. The same study also revealed that this approach can be also used to control the cellular localization and biological activity of proteins in live cells [28].

The split DnaE intein has been also used *in vivo* for the C-terminal labeling of a pleckstrin homology (PH) domain with quantum dots inside *Xenopus* embryos [53]. The same authors have also reported the use of the a split mini-DnaB intein for N-terminal labeling of proteins *in vivo* [54] therefore opening the possibility for multiprobe labeling using orthogonal split-inteins.

The regioselective labeling of unique Cys residues in proteins with multiple Cys amino acids can be also accomplished using a combination of cysteine bioconjugation and PTS [55]. For example, in the presence of more than one cysteine, a short cysteine-containing peptide fused to I_C intein fragment and modified at its sulfhydryl group can be employed for the labeling of proteins by trans-splicing with a POI fused to the I_N intein fragment. This approach makes possible to chemoselective labeling a protein containing multiple cysteines. The artificially split *Mxe* GyrA intein and the *Ssp* DnaB inteins have been employed for the labeling of several proteins using this approach [55]. Alternatively, a tagged peptide containing a cysteine can first be labeled and later attached to the protein of interest, leaving reactive residues in the protein of interest unaffected. These approaches allow labeling of proteins by site-specific Cys bioconjugation with commercially available cysteine modifying probes.

3.3 Segmental isotopic labeling

Segmental isotopic labeling is a technique that allows for the selective labeling of a protein segment with stable isotopes having unique spectroscopic properties (Fig. 5B). For example,

this technique permits the generation of segmental isotopically labeled proteins with NMR-active isotopes such as ^{13}C and ^{15}N , which can help to overcome some of the size limitations associated with the structural analysis of proteins by heteronuclear NMR [24, 56]. Although the NMR signals from large proteins can be detected they show significant spectral overlap which is difficult to resolve even in 3D- or 4D-NMR spectra [57]. Decreasing such spectral complexity can be accomplished by generating protein samples where only selected protein segments or amino acids are labeled with NMR active nuclei, therefore removing the signals coming from the rest of the unlabeled molecule. Both intein-based techniques, EPL and PTS, have been successfully used to generate proteins segmentally isotopic labeled for NMR studies [24, 56].

EPL was used to produce a segmental isotopic labeled *Mxe* GyrA intein, where the C-terminal N-extein residue was labeled with ^{13}C and ^{15}N , in order to study more in detail the mechanism of protein splicing [58]. This allowed for the first time to measure $^1\text{J}_{\text{NC}}$ coupling constant of the amide bond at the N-extein-intein junction, indicating that this peptide bond was highly polarized thus strongly supporting the idea that the first step in protein splicing is facilitated, in part, by the destabilization of the peptide bond at the N-extein-intein junction.

As mentioned earlier, EPL makes possible to ligate a uniformly labeled protein fragment to the rest of the unlabeled protein therefore significantly decreasing the spectral complexity of the sample, which allows the analysis of the signals from the labeled fragment in the context of the whole protein. By using an optimized on column EPL protocol in combination with transverse relaxation-optimized NMR spectroscopy (TROSY), it was possible solve the solution structure of the two C-terminal RNA recognition motifs (RRM3 and RRM4) of the polypyrimidine tract binding protein (PTB) in the context of the whole protein, which contains about 700 residues [59]. The final structure showed a rather large interdomain interface, which resulted in the a two RRM domains adopting a very unusual positioning relative to one another. Based on these findings, the authors hypothesized that this unusual structure is responsible for inducing the formation of RNA loops, which could prevent splicing by sequestering either a short alternative-exon or a branch point within these RNA loops [59].

A similar on column procedure has been also reported for the preparation of segmentally labeled constructs of the lipoprotein apolipoprotein E3 (apoE) [34]. The NMR spectrum of apoE is significantly overlapped as is a 299-residue helical protein. Reduction of the NMR spectral complexity for a complete spectral assignment was accomplished by producing several segmental labeled apoEs, in which one domain was ^{13}C , ^{15}N -labeled whereas the other domains were deuterated [34].

Split-intein-mediated PTS has been also used for this task and provides an alternative to EPL for the production segmental labeled protein samples [56b]. This can be accomplished by using two separately purified protein fragments fused to the corresponding split-intein fragments prepared with different isotopic enrichments that are then ligated *in vitro* (Fig. 5B). Initially, PTS segmental isotopic labeling was performed using artificial split-inteins, but this reaction required refolding of the protein fragments [56a]. The use of natural occurring split inteins avoids the need for refolding, making it a more efficient process [60].

Several split inteins have been used including native split *Ssp* DnaE [60a] and *Npu* DnaE [61], and the artificially split PI-*PfuI* [62] and PI-*PfuII* [63].

PTS can also be used for segmental isotopic protein labeling in live cells [27a]. This can be accomplished by the sequential expression of the labeled and unlabeled fragments in medium containing different isotopes, respectively, using a compatible dual expression system with different inducible promoters (Fig. 5B). For example, in-cell PTS was used to fuse unlabeled immunoglobulin-binding protein G domain B1 (GB1) to the C-terminus of the yeast prion protein Sup35 (residues 1–61). This approach yielded a Sup35 fusion protein with improved stability and solubility provided by the NMR-invisible GB1-tag [27a].

PTS has also expanded the use of solid state NMR spectroscopy (ssNMR) [64]. The use of ssNMR provides a unique approach for the structural analysis of large protein assemblies, which are usually challenging to study due to their non-crystalline form and poor solubility. The limited resolution and spectral crowding are also currently major limitations found when using ssNMR. The use of segmental isotope labeling provides a powerful approach to overcome these limitations as in ssNMR the line widths of the signals are independent on the size of the proteins being analyzed. In this context, segmental isotope labeling provides a scalable approach to reduce spectral complexity aiding in the assignment of repetitive amino acid sequences. In addition, it can also be easily combined with other selective isotope labeling strategies [64]. In a recent report, the highly efficient *Npu* DnaE split-intein was used to study the assembly of HIV-1 capsomere proteins [61a] and the aggregation process of the yeast prion Sup35NM [61b].

The use of isotopically labeled proteins has also extended the structural analysis of proteins by vibrational spectroscopy. For example, EPL was used for the preparation of a segmental ¹³C-labeled version of phospholipase A2 (PLA₂) that in combination with polarized attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy allowed to determine the mode of interaction of PLA₂ with membranes [65]. By using different segmental labeled versions of PLA₂, where only two of the three α -helices of this protein were ¹³C-labeled, the authors generated enough structural constraints to establish a structural model for the membrane-bound PLA₂ [65].

3.4 Polypeptide backbone cyclization

Backbone cyclization of recombinant polypeptides can be achieved both *in vitro* and *in vivo* by either EPL or PTS (see ref. [66] for more detailed reviews on this topic).

Backbone cyclization of a protein using EPL can be accomplished by placing a cysteine at the N-terminus of the target protein while the C-terminus is fused to an N-terminus of a modified Cys intein engineered to favor N-terminal cleavage. A cysteine can either be generated by introducing an upstream intein or by conventional proteolytic cleavage. The Cys residue can then react in an intramolecular fashion with an α -thioester generated by the downstream intein, thus providing a backbone cyclized polypeptide (Fig. 6).

The first use of backbone cyclization of a protein using EPL both *in vitro* and in cell was reported on the SH3 domain of c-Crk protein adaptor [67]. Cyclization of the SH3 domain

resulted in faster protein folding with increased thermal stability and improved biological activity [67b, 68]. The same approach has been used to cyclize other with similar results on stability and activity, including green fluorescent protein (GFP) and β -lactamase [69].

The use of EPL to produce backbone-cyclized polypeptides inside live cells allows the production and genetic selection of complex libraries of cyclic polypeptides using standard recombinant techniques. The use of highly stable cyclic scaffolds provides biologists with a powerful molecular tool that can be used to decipher genomic information encoding protein interaction networks and cellular biochemical pathways [70]. For example, in-cell EPL-mediated backbone cyclization has been used to generate cyclotide-based libraries using standard heterologous expression systems [66b, 71]. Cyclotides are microproteins with a head-to-tail cyclized backbone, that are further stabilized by the presence of three disulfide bonds forming a cysteine knot. The cysteine knot motif (CCK) provides cyclotides exceptional resistance to chemical, physical and biological degradation, making them excellent molecular scaffolds for the development of novel peptide-based therapeutics and bioimaging agents [72]. A similar approach has also been used for in cell production of naturally-occurring disulfide-rich backbone cyclized peptides SFTI-1 [73] and the θ -defensins RTD-1 [74] and HTD-2 [75].

PTS provides an alternative for the production of backbone cyclized polypeptides [66c, 76]. Rearranging the order of the intein fragments, i.e. by fusing the I_N and I_C fragments to the C- and N-termini of the polypeptide to cyclize, allows the trans-splicing reaction to produce a backbone-cyclized polypeptide (Fig. 7). This approach was first reported by using the *Ssp* DnaE split-intein for the generation of small backbone-cyclized polypeptides as well as large genetically-encoded library of small cyclic peptides in bacterial cells [76–77]. The production of backbone-cyclized or circular proteins has been also accomplished by using PTS [66a, 78]. The artificially split *Ssp* DnaB mini-intein was employed for the production of a backbone-cyclized form of the TEM-1 β -lactamase in the bacterial periplasm [79].

It is worth noting that many split inteins, like for example *Ssp* DnaE or *Ssp* DnaB, require the presence of specific amino acids at both extein-intein junctions for efficient protein splicing to happen [77]. Using the *Npu* DnaE split-intein seems to overcome this limitation. This naturally-split intein is one of the most efficient split inteins reported thus far with reported rates of protein *trans*-splicing ($\tau_{1/2} \sim 60$ s), a high splicing efficiency ($\sim 80\%$), and more importantly high sequence tolerance at the extein-intein junctions allowing traceless cyclization [78b, 80]. This intein has been widely used for the efficient production of native disulfide-rich backbone cyclic polypeptides such as cyclotides in different heterologous expression systems [78]. Other disulfide-rich cyclic polypeptides including trypsin inhibitor SFTI-1 and several θ -defensins have also been produced in high yield in *E. coli* expression systems using this approach [80]. In addition, a genetically-encoded RTD-1-based peptide library with a complexity of $\approx 10^7$ was recently produced in live *E. coli* cells [80b]. The possibility of generating genetically-encoded RTD-1-based peptide libraries in live *E. coli* cells in combination with the use of appropriate in-cell reporters should allow the development of directed evolution technologies using the cyclic peptide RTD-1 as a molecular scaffold.

PTS has also used for the expression of folded cyclotides in eukaryotic microorganisms such as yeast *S. cerevisiae* [78a]. In this work, PTS was used for the in-cell production of a novel bioactive cyclotide able to inhibit α -synuclein-induced cytotoxicity in yeast cells. This allowed for the first time to perform a phenotypic screen for the selection of cyclotide sequences with biological activity in a yeast synucleopathy model, therefore opening the possibility to achieve in-cell phenotypic screening of cyclotide-based libraries against other cellular pathologies. In addition, employing eukaryotic microorganisms for screening purposes provides a more biologically relevant cellular background for phenotypic screening and facilitates the production of cyclotide-based libraries containing post-translational modifications, such as phosphorylation and/or glycosylation, that are not available when using bacterial expression systems.

4. Molecular biosensors

Intein-based biosensors are designed for sensing molecular interactions in complex biological systems such as live cells. They have been used to observe molecular mechanisms such as protein-protein interactions [81], and protein translocation into the nucleus [82], endoplasmic reticulum (ER) [83] and mitochondria [84]. Intein-based biosensors have also been used for the detection of small molecules [85], redox state [10, 86] and protease activity [87].

Intein-based biosensors are made from a combination of individual protein domains. They are usually comprised of three distinct modules, the sensing and reporter modules, and an intein-derived signal transduction module. Intein-based biosensors can be genetically-encoded and therefore can be used inside live cells. Once expressed inside a live cell they are activated when a signal triggers its sensory module, which induces intein splicing resulting in a change in the reporter activity (Fig. 8). Many of these biosensors use conditional protein splicing, which uses an external *stimulus* to activate its protein splicing activity. Binding of proteins or small molecules, a change in pH, light, temperature, or redox state can be used to activate conditional protein splicing [88]. Protein splicing of the biosensor restores the activity of its reporter domain, producing a detectable and sometimes quantifiable signal.

4.1 Biosensors for protein-protein interactions

Intein-based biosensors are widely used to detect protein-protein interactions. These protein-protein biosensors use PTS facilitated by artificial or engineered split inteins with low affinity for each other [17a, 26]. These types of biosensors are comprised of two fusion proteins; each containing a split intein fragment (I_C or I_N), a protein binding site or protein domain and a portion of a split reporter protein. The interacting proteins or protein domains are linked to the N- and C-termini of the I_C and I_N fragments of a split intein, respectively. These constructs are also linked to the segments of an appropriate split reporter module (Fig. 8A). When an interaction between the two interacting proteins or protein domains occurs, the split inteins fragments are brought into close proximity facilitating the reconstitution of a fully functional intein *in trans*, which leads to the production of the full-length reporter protein and the corresponding associated signal (Fig. 8A). Reporters based on the green fluorescent protein (GFP) and luciferase enzyme have been used to detect and quantify

protein-protein interactions in systems ranging from single cells to transgenic animals [16e, 29a, 29b, 89]. For example, this approach has been used to measure the interaction between calmodulin and its target peptide M13 using an artificially split *Sce* VMA1 intein GFP-based biosensor in *E. coli* [89]. A luciferase-based biosensor using the natural split *Ssp* DnaE intein was also employed to detect the interaction between phosphorylated insulin receptor substrate 1 and the SH2 domain of PI3-Kinase in live mammalian cells [81a]. A similar approach using a firefly luciferase-based biosensor was used for non-invasive live imaging in mice of the protein-protein interaction between the MyoD and Id [81d]. The same firefly luciferase reporter has also been used for the production of new reporter gene assay for the detection of cytosolic protein-protein interactions using the protein splicing of transcription factors with DnaE split-inteins. This intein-based reporter used epidermal growth factor (EGF)-induced Ras-Raf1 binding interactions as a model system, to activate splicing of transcription factors, therefore activating firefly luciferase transcription in mammalian cells [81c]. This signal-amplification approach allows to improve the sensitivity of this type of intein-based biosensors.

4.2 Biosensors for protein-DNA interactions

The same biosensor concept has been also used to detect changes in the DNA methylation patterns in living cells by using bioluminescence [90]. In this case, the biosensor is composed of two chimeric proteins, where each one contains a specific polydactyl Zn-finger domain fused to the corresponding split-intein and a split-luciferase fragments.

When the two polydactyl Zn-finger domains bind their DNA targets, the intein splicing activity is initiated, resulting in the reconstitution of the luciferase reporter. This artificial Zn-finger domain was designed to target the DNA sequences near the promotor region of L1PA2 subfamily of Line-1 retroelements (Fig. 8B). Therefore, when the biosensor binds to its target DNA sequence a luminescent signal is produced thereby allowing the detection in loss of epigenetic silencing and increased DNA accessibility of retroelements. This biosensor allowed the authors to study the response of demethylating therapeutic agents in human cells [90].

4.3 Biosensors for protein localization and internalization

Intein biosensors can be also used to detect protein localization in living cells using a similar approach. In order to allow protein splicing to only occur in a particular selected cellular compartment, these biosensors are designed to have one of the fusion proteins used on the intein-based biosensor tagged with an appropriate organelle-specific targeting sequence (Fig. 8C) [81d, 82–84, 90].

The *Ssp* DnaE intein in combination with a split *Renilla* luciferase (RLuc) and a glucocorticoid receptor (GR) has been used to detect the levels of corticosterone in mice [91]. This was accomplished by fusing the GR with the C-terminal halves of RLuc and DnaE, which localizes in the cytosol. A second fusion protein containing the N-terminal halves of RLuc and DnaE was localized in the nucleus by adding a nuclear localization signal. Once corticosterone induces GR translocation into the nucleus, the C-terminal RLuc interacts with the N-terminal fusion protein located in the nucleus therefore reconstituting

full-length by protein splicing. This approach was shown to provide a quantitative bioluminescence assay of the determine the amount of GR translocated into the nucleus.

The use of intein-based biosensors has been also employed to observe cellular internalization of G-protein-coupled receptors (GPCRs) [92]. GPCR internalization is dependent on the interaction between the corresponding activated GPCR and β -arrestin 2. The biosensor in this case used the *Npu* DnaE intein in combination with split RLuc to detect internalization through the binding of activated GPCR and β -arrestin2 [92a]. A similar biosensor approach using split versions of both RLuc or firefly luciferase was also recently employed to detect the interaction between activated GPCRs with Rab5 or β -arrestin2 [92b].

4.4 Biosensors for protease activity

Biosensors based on split inteins have also employed to detect protease activity. The *Ssp* DnaE intein in combination with a split version of the firefly luciferase was used for the quantitative detection of caspase-3 activity in live mice on real time [87]. The biosensor was based on a backbone cyclized luciferase protein reporter containing the caspase-3 recognition sequence. In this case the *Ssp* DnaE split intein was used to cyclize the luciferase reporter. When cyclized, the luciferase reporter was enzymatically inactive due to the steric hindrance imposed by the backbone cyclization. In the presence the of caspase-3, cleavage of the cyclized biosensor restored the activity of the luciferase, allowing real-time detection of caspase *in vivo* (Fig. 8D). A similar design has also been used to detect the protease activity of the hepatitis A virus 3C protease in human cells although in this case the more efficient *Npu* DnaE split-intein was used to produce the backbone cyclized version of the reporter enzyme [93].

More recently caged intein 'zymogens' have been proposed as biosensors for proteolytic activity [94]. In this case, the sensor is designed around a caged version of the *Npu* DnaE split-intein that can be uncaged by a specific protease to produce a fully functional split-intein able to splice *in trans* to produce the desired protein reporter. This system was tested using a TEV-activated split GFP intein zymogen system that was able to produce GFP in the presence of the TEV protease in live cells [94].

4.5 Biosensors for detecting redox state

Since many inteins splice using cysteine-based chemistry, their activity can be modulated by the redox state through disulfide bond formation. This can be accomplished by reversible trapping of the intein catalytic cysteine/s residue/s through a disulfide bond [86]. The first example reported was used to detect redox conditions *in vitro* and inside live *E. coli* cells. The sensor used an engineered version of the *Ssp* DnaE intein with a disulfide bond formed between of the N-terminal intein Cys and an additional Cys at the C-terminal region of N-extein segment. This intein was fused to two fluorescent proteins allowing detection of activity through changes in the FRET signal. In the inactive state, the disulfide bond is undisturbed resulting in high FRET levels. When the redox state of the environment allows the reduction of the disulfide, the intein is activated severing the fluorescent proteins and resulting in loss of the FRET signal (Fig. 8E).

4.6 Biosensors for detecting small molecules

Intein-mediated biosensors have been used to detect small molecules both *in vitro* and *in vivo*. Early work showed that inteins can be rationally designed for regulating protein splicing *in vivo* in the presence of small molecules such as human thyroid hormone (Fig. 8F) [85a, 85b]. In this work an artificial mini-intein originally derived from the *Mtu* RecA was used by replacing its entire central endonuclease domain with the ligand-binding domain of the human thyroid hormone receptor (TR). TR is a member of the nuclear receptor superfamily that has been shown to undergo a conformational change upon thyroid hormone binding that results in bringing its N- and C-termini in relative proximity as well as a stability enhancement of the protein complex. Inserting the TR binding domain in one of the loops of the intein abolished its splicing activity, which could be later restored by adding the thyroid hormone or synthetic analogs. This allosteric intein design was employed for the conditional activations of a variety of different proteins in *E. coli* in a dose-dependent manner. The same authors also used a combination of genetic selection and directed evolution to produce an additional controllable intein whose splicing activity could be inhibited by the presence of an orthogonal set of synthetic estrogen ligands [85c].

Another allosteric intein biosensor using the same design principle was used to detect isoform-selective ligands for human thyroid receptors α -1 and β -1 (TR α -1 and TR β -1) [95]. This allosteric intein-based biosensor combines the structure of a hormone receptor ligand-binding domain with a thymidylate synthase reporter enzyme as reporter. When expressed in *E. coli* allowed to have a hormone-dependent growth phenotype, thereby allowing to detect potentially therapeutic subtype-selective agonistic compounds when combined with other intein-based receptors (e.g. TR α -1 *versus* TR β -1) [95].

A similar approach has been employed to detect estrogen levels using a β -galactosidase reporter enzyme [96]. The sensor protein was constructed by inserting an estrogen-sensitive *Sce* VMA intein into the LacZ gene. This engineered intein was made by replacing the endonuclease region of *Sce* VMA intein with the estrogen binding domain of the human estrogen receptor- α (ER- α). In the presence of estrogen, the splicing of the engineered intein was induced resulting in the production of LacZ protein.

5. Gene delivery and control in transgenic organisms

Split-inteins have been employed to control the heterologous expression of genes in transgenic organisms (see ref. [17a, 17b] for recent reviews on this topic). This approach relies on splitting of the target protein to be transferred into two segments that can be post-translationally reconstituted *in vivo* by PTS. One of the advantages of this approach is that it minimizes the risk of transferring the particular gene to undesired hosts, since to be effective the transfer of the whole gene would require the transfer of both gene fragments separately. For example, the bacterial 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) can be reconstituted by PTS in the chloroplasts of *Nicotiana tabacum* for the generation of an herbicide-resistant transgenic crop [97]. In this approach the split gene fragments were introduced into nuclear and chloroplast chromosomes significantly inhibiting the potential for recombination and transfer to an undesired host. Similar approaches have also been employed for the reconstitution of biologically active β -glucuronidase in the plant

Arabidopsis thaliana [98] as well as in the leaves of other plants, including soybean and barley among others [99].

The thermostable self-splicing bacterial DnaE-1 intein from *Thermus thermophilus* was recently employed to control the heterologous production of a thermostable xylanase (XynB) from *Dictyoglomus thermophilum* in maize [100]. The use of random mutagenesis via error-prone PCR allowed the generation of xylanase-intein protein mutants able to splice only at high temperatures and to produce wild-type active xylanase after splicing, while preventing the catalytic activity of the unprocessed xylanase-intein protein fusion. Expression of the XynB enzyme by the corn plant allows the production of a “self-processing” corn able to hydrolyze its own cellulosic biomass to simple sugars for fermentation. Transgenic maize constitutively expressing active XynB exhibits shriveled seeds and low fertility, however, transgenic maize expressing the thermoregulated xylanase-intein fusion produces normal seeds. Treatment of this transgenic maize with high temperature restores wild-type xylanase activity, which allows the pretreatment of the plant cellulosic biomass to release soluble sugars.

A split variant of the DnaB intein was employed to facilitate allow the polymerization of spider silk flagelliform protein in the leaves of tobacco plant [101]. The repeated ligation mediated by PTS of the corresponding FLAG monomers was able to produce large multimers. This approach circumvents the need for using highly repetitive transgenes, that could lead to genetic and transcriptional instability of the corresponding gene.

PTS has also made possible the delivery of large genes in animal cells. For example, the split DnaB mini-intein has been used to produce *in vitro* and *in vivo* the heavy and light chains of the B-domain deleted factor VIII by using two separate viral vectors [102]. These results prove that the production of proteins that are too large to be delivered using a single viral vector can be efficiently accomplished using PTS.

The conditional production of Cre recombinase in mice has been also accomplished using the split DnaE intein [103]. In this work, the authors used two different promoters to control the expression of the two split intein-Cre constructs, which allowed to interrogate the expression patterns of two genes or promoters while allowing the production of proteins under control of the Cre-Lox P system [103].

More recently, an intein-mediated split-Cas9 system has been developed to incorporate Cas9 into cells and reconstitute nuclease activity efficiently [104]. The use of CRISPR/Cas9 makes possible to target virtually any gene in any organism. However, one of the major limitations to its application in gene therapy is the size of the Cas9 gene (>4 kb), which makes difficult its efficient delivery using an adeno-associated virus (rAAV) delivery approach. The use of the split *Npu* DnaE-based PTS allowed the authors to develop a split-Cas9 system able to reconstitute full length Cas9 from two smaller genes thereby bypassing the packaging limit associated with the use of rAAVs for gene delivery. In this approach, the two halves of Cas9 were fused to the I_N and I_C fragments of the *Npu* DnaE intein that after co-expression were able to produce full-length Cas9 by PTS [104]. The same authors used the same approach to design a split-Cas9D10A nickase version. These exciting results

should allow the use of AAV-based technologies for gene delivery to treat a wider variety of diseases for which therapeutic trans-genes exceed the packaging limitation of present AAV vectors [104].

6. Concluding remarks

Since the discovery of protein splicing in the early 1990s, the use of protein splicing domains or inteins has become more commonly used in laboratories worldwide. In many cases, the biotechnological applications of protein splicing have started to move from proof-of-concept experiments to productive applications. A key aspect for the large success on the use of inteins for many of the biotechnological applications have been the discovery of ultrafast and efficient inteins [105] and the identification of split-inteins with large sequence variation tolerance on the intein-extein junctions like for example *Npu* DnaE-based split inteins [80b, 106]. The use of such inteins for example allows the rapid creation of large genetically-encoded libraries of cyclic peptides without the need to incorporate the native extein sequences [80b]. Extremely efficient inteins should also allow the labeling of proteins and produce biosensors that can work on real time on live cells and/or organisms [28, 53–54].

Despite the success on using inteins for many applications, there is still a need to have a detailed understanding on how the intein fold coordinates the chemistry of its multiple active sites. The findings indicating that intein activity can be modulated by the introduction of mutations located far away from the active site seems to indicate that the allosteric networks found in the intein molecular framework could play a larger role in regulating protein splicing activity than previously anticipated. A more detailed understanding of the dynamics and structure of inteins should provide information essential to engineer more efficient protein splicing domains as better biotechnology tools.

In summary, the ability to create peptide bonds using an intein-mediated protein splicing has provided a tremendous tool for many biological and biotechnological applications. In principle, there is no limit to the usability of inteins in future applications.

Acknowledgements

This work was supported by National Institutes of Health Research Grant R01-GM113363 (J.A.C.), Department of Defense Congressionally Directed Medical Research Programs in Lung Cancer Grant LC150051 (J.A.C), BROAD Medical Research Program-Crohn's & Colitis Foundation of America Grant #483566 (J.A.C.), Lupus Research Institute (J.A.C) and Whittier Foundation (J.A.C).

References

- [1]. Perler FB; Davis EO; Dean GE; Gimble FS; Jack WE; Neff N; Noren CJ; Thorner J; Belfort M Nucleic Acids Res. 1994, 22 (7), 1125–1127. [PubMed: 8165123]
- [2]. (a) Mathys S; Evans TC; Chute IC; Wu H; Chong S; Benner J; Liu X-Q; Xu M-Q Gene 1999, 231 (1), 1–13; [PubMed: 10231563] (b) Myscofski DM; Dutton EK; Cantor E; Zhang A; Hruby DE Prep. Biochem. Biotechnol 2001, 31 (3), 275–290; [PubMed: 11513092] (c) Wood DW; Wu W; Belfort G; Derbyshire V; Belfort M Nat. Biotechnol 1999, 17, 889–892. [PubMed: 10471931]
- [3]. (a) Hirata R; Ohsumi Y; Nakano A; Kawasaki H; Suzuki K; Anraku YJ Biol. Chem 1990, 265 (12), 6726–6733; (b) Kane PM; Yamashiro CT; Wolczyk DF; Neff N; Goebel M; Stevens TH Science 1990, 250 (4981), 651–7. [PubMed: 2146742]

- [4]. Perler FB *Nucleic Acids Res.* 2002, 30 (1), 383–4. [PubMed: 11752343]
- [5]. (a)Gimble FS; Thorner J *Nature* 1992, 357 (6376), 301; [PubMed: 1534148] (b)Belfort M; Roberts RJ *Nucleic Acids Res.* 1997, 25 (17), 3379–3388; [PubMed: 9254693] (c)Barzel A; Naor A; Privman E; Kupiec M; Gophna U *Biochem. Soc. Trans* 2011, 39, 169–173. [PubMed: 21265767]
- [6]. Hafez M; Hausner G *Genome* 2012, 55 (8), 553–569. [PubMed: 22891613]
- [7]. (a)Pavankumar TL *Microorganisms* 2018, 6 (1);(b)Petrokovski S *Trends Genet.* 2001, 17 (8), 465–472. [PubMed: 11485819]
- [8]. Dalgaard JZ; Moser MJ; Hughey R; Mian IS J. *Comput. Biol* 1997, 4 (2), 193–214. [PubMed: 9228618]
- [9]. (a)Frischkorn K; Sander P; Scholz M; Teschner K; Prammananan T; Böttger EC *Mol. Microbiol* 1998, 29 (5), 1203–1214; [PubMed: 9767588] (b)Papavinasundaram K; Colston MJ; Davis EO *Mol. Microbiol* 1998, 30 (3), 525–534. [PubMed: 9822818]
- [10]. Callahan BP; Topilina NI; Stanger MJ; Van Roey P; Belfort M *Nat. Struct. Mol. Biol* 2011, 18 (5), 630–633. [PubMed: 21460844]
- [11]. Derbyshire V; Wood DW; Wu W; Dansereau JT; Dalgaard JZ; Belfort M *Proc. Natl. Acad. Sci. USA* 1997, 94 (21), 11466–71. [PubMed: 9326633]
- [12]. Wu H; Hu Z; Liu XQ *Proc. Natl. Acad. Sci. USA* 1998, 95, 9226–9231. [PubMed: 9689062]
- [13]. Caspi J; Amitai G; Belenkiy O; Petrokovski S *Mol. Microbiol* 2003, 50 (5), 1569–77. [PubMed: 14651639]
- [14]. Iwai H; Züger S; Jin J; Tam P-H *FEBS Lett.* 2006, 580 (7), 1853–1858. [PubMed: 16516207]
- [15]. Debelouchina GT; Muir TW Q. *Rev. Biophys* 2017, 50, e7. [PubMed: 29233219]
- [16]. (a)Appleby JH; Zhou K; Volkmann G; Liu X-Q *Biol. Chem* 2009, 284 (10), 6194–6199; (b)Aranko AS; Züger S; Buchinger E; Iwai H *PLoS One* 2009, 4 (4), e5185; [PubMed: 19365564] (c)Lee Y-T; Su T-H; Lo W-C; Lyu P-C; Sue S-C *PLoS One* 2012, 7 (8), e43820; [PubMed: 22937103] (d)Sun W; Yang J; Liu X-Q *Biol. Chem* 2004, 279 (34), 35281–35286; (e)Volkmann G; Liu X-Q *PLoS One* 2009, 4 (12), e8381. [PubMed: 20027230]
- [17]. (a)Wood DW; Camarero JA J. *Biol. Chem* 2014, 289 (21), 14512–9; [PubMed: 24700459] (b)Topilina NI; Mills KV *Mob. DNA* 2014, 5 (1), 5; [PubMed: 24490831] (c)Belfort M *Curr. Opin. Microbiol* 2017, 38, 51–58. [PubMed: 28482231]
- [18]. Volkmann G; Mootz HD *Cell. Mol. Life Sci* 2013, 70 (7), 1185–1206. [PubMed: 22926412]
- [19]. Petrokovski S *Protein Sci.* 1994, 3 (12), 2340–2350. [PubMed: 7756989]
- [20]. Frutos S; Goger M; Giovani B; Cowburn D; Muir TW *Nat. Chem. Biol* 2010, 6 (7), 527–533. [PubMed: 20495572]
- [21]. Southworth MW; Benner J; Perler FB *EMBO J.* 2000, 19 (18), 5019–26. [PubMed: 10990465]
- [22]. Amitai G; Dassa B; Petrokovski SJ *Biol. Chem* 2004, 279 (5), 3121–31.
- [23]. (a)Southworth MW; Amaya K; Evans TC; Xu M-Q; Perler FB *BioTechniques* 1999, 27 (1), 110–4, 116, 118–20; [PubMed: 10407673] (b)Chong S; Montenegro GE; Zhang A; Cantor EJ; Liao W; Xu M-Q; Benner J *Nucleic Acid Res.* 1998, 26 (22), 5109–5115; [PubMed: 9801307] (c)Chong S; Mersha FB; Comb DG; Scott ME; Landry D; Vence LM; Perler FB; Benner J; Kucera RB; Hirvonen CA; Pelletier JJ; Paulus H; Xu MQ *Gene* 1997, 192 (2), 271–81. [PubMed: 9224900]
- [24]. Berrade L; Camarero JA *Cell. Mol. Life Sci* 2009, 66 (24), 3909–22. [PubMed: 19685006]
- [25]. Camarero JA; Muir TW *Curr. Protoc. Protein Sci* 1999, (18.4), 1–21.
- [26]. Li YF *Biotechnol. Lett* 2015, 37 (11), 2121–2137. [PubMed: 26153348]
- [27]. (a)Iwai H; Züger S *Biotechnol. Genet. Eng. Rev* 2007, 24 (1), 129–146; [PubMed: 18059630] (b)Wilkinson KD; Gan-Erdene T; Kolli N *Methods Enzymol.* 2005, 399, 37–51; [PubMed: 16338347] (c)Muir TW; Sondhi D; Cole PA *Proc. Natl. Acad. Sci. U S A* 1998, 95 (12), 6705–10; [PubMed: 9618476] (d)Cotton GJ; Ayers B; Xu R; Muir TW *J. Am. Chem. Soc* 1999, 121 (5), 1100–1101.
- [28]. Borra R; Dong D; Elnagar AY; Woldemariam GA; Camarero JA J. *Am. Chem. Soc* 2012, 134 (14), 6344. [PubMed: 22404648]
- [29]. (a)Kanno A; Ozawa T; Umezawa Y *Methods Mol. Biol* 2011, 251–258;(b)Kanno A; Ozawa T; Umezawa Y *Methods Mol. Biol* 2009, 185–192;(c)Kanno A; Umezawa Y; Ozawa T *Methods Mol. Biol* 2009, 105–114.

- [30]. Wood DW; Harcum SW; Belfort G Industrial applications of intein technology In Homing endonucleases and inteins, B. M; W. DW; S. BL; D. V, Eds. Springer: Berlin, Heidelberg, 2005; Vol. 16, pp. 345–364.
- [31]. Shah NH; Muir TW Chem. Sci 2014, 5 (2), 446–461. [PubMed: 24634716]
- [32]. Chong S; Shao Y; Paulus H; Benner J; Perler FB; Xu MQ J. Biol. Chem 1996, 271 (36), 22159–68. [PubMed: 8703028]
- [33]. (a)Evans TC; Benner J; Xu MQ Protein Sci. 1998, 7 (11), 2256–2264; [PubMed: 9827992] (b)Severinov K; Muir TW J. Biol. Chem 1998, 273 (26), 16205–9. [PubMed: 9632677]
- [34]. Zhao W; Zhang Y; Cui C; Li Q; Wang J Protein Sci. 2008, 17 (4), 736–47. [PubMed: 18305193]
- [35]. Xu M-Q; Evans TC Methods 2001, 24 (3), 257–277. [PubMed: 11403575]
- [36]. Wood DW; Derbyshire V; Wu W; Chartrain M; Belfort M; Belfort G Biotechnol. Prog 2000, 16 (6), 1055–63. [PubMed: 11101334]
- [37]. (a)Xie YG; Luan C; Zhang HW; Han FF; Feng J; Choi YJ; Groleau D; Wang YZ Protein Pept. Lett 2013, 20 (1), 54–60; [PubMed: 22670762] (b)Wang Z; Li N; Wang Y; Wu Y; Mu T; Zheng Y; Huang L; Fang X Protein Expr. Purif 2012, 82 (1), 174–8. [PubMed: 22178731]
- [38]. (a)Banki MR; Gerngross TU; Wood DW Protein Sci. 2005, 14 (6), 1387–1395; [PubMed: 15883185] (b)Wu W-Y; Mee C; Califano F; Banki R; Wood DW Nat. Protoc 2006, 1 (5), 2257–2262; [PubMed: 17406465] (c)Ge X; Yang DS; Trabbic-Carlson K; Kim B; Chilkoti A; Filipe CD J. Am. Chem. Soc 2005, 127 (32), 11228–11229. [PubMed: 16089436]
- [39]. (a)Floss DM; Schallau K; Rose-John S; Conrad U; Scheller J Trends Biotechnol. 2010, 28 (1), 37–45; [PubMed: 19897265] (b)Fong BA; Wu W-Y; Wood DW Protein Expression Purif. 2009, 66 (2), 198–202.
- [40]. Wieczorek R; Pries A; Steinbüchel A; Mayer FJ Bacteriol. 1995, 177 (9), 2425–2435.
- [41]. Georgiou G; Jeong KJ Protein Sci. 2005, 14 (6), 1385–1386. [PubMed: 15929993]
- [42]. Lu W; Sun Z; Tang Y; Chen J; Tang F; Zhang J; Liu J-N J. Chromatogr. A 2011, 1218 (18), 2553–2560. [PubMed: 21397239]
- [43]. Berrade L; Kwon Y; Camarero JA ChemBioChem 2010, 11 (10), 1368–72. [PubMed: 20512791]
- [44]. Guan D; Ramirez M; Chen Z Biotechnol. Bioeng 2013, 110 (9), 2471–2481. [PubMed: 23568256]
- [45]. Shi C; Meng Q; Wood DW Appl. Microbiol. Biotechnol 2013, 97 (2), 829–835. [PubMed: 23212673]
- [46]. Chacko BM; Qin BY; Tiwari A; Shi G; Lam S; Hayward LJ; De Caestecker M; Lin K Mol. Cell 2004, 15 (5), 813–23. [PubMed: 15350224]
- [47]. Qin BY; Lam SS; Correia JJ; Lin K Genes Dev. 2002, 16 (15), 1950–63. [PubMed: 12154125]
- [48]. (a)Durek T; Alexandrov K; Goody RS; Hildebrand A; Heinemann I; Waldmann HJ Am. Chem. Soc 2004, 126 (50), 16368–78; (b)Brunsveld L; Watzke A; Durek T; Alexandrov K; Goody RS; Waldmann H Chemistry (Easton) 2005, 11 (9), 2756–72.
- [49]. (a)Rak A; Pylypenko O; Durek T; Watzke A; Kushnir S; Brunsveld L; Waldmann H; Goody RS; Alexandrov K Science 2003, 302 (5645), 646–50; [PubMed: 14576435] (b)Pylypenko O; Rak A; Durek T; Kushnir S; Dursina BE; Thomae NH; Constantinescu AT; Brunsveld L; Watzke A; Waldmann H; Goody RS; Alexandrov K EMBO J. 2006, 25 (1), 13–23; [PubMed: 16395334] (c)Guo Z; Wu YW; Das D; Delon C; Cramer J; Yu S; Thuns S; Lupilova N; Waldmann H; Brunsveld L; Goody RS; Alexandrov K; Blankenfeldt W EMBO J. 2008, 27 (18), 2444–56. [PubMed: 18756270]
- [50]. Flavell RR; Muir TW Acc. Chem. Res 2009, 42 (1), 107–16. [PubMed: 18939858]
- [51]. (a)Shogren-Knaak MA; Fry CJ; Peterson CL J. Biol. Chem 2003, 278 (18), 15744–8; [PubMed: 12595522] (b)Shogren-Knaak M; Ishii H; Sun JM; Pazin MJ; Davie JR; Peterson CL Science 2006, 311 (5762), 844–7; [PubMed: 16469925] (c)He S; Bauman D; Davis JS; Loyola A; Nishioka K; Gronlund JL; Reinberg D; Meng F; Kelleher N; McCafferty DG Proc. Natl. Acad. Sci. U. S. A 2003, 100 (21), 12033–8; [PubMed: 14530408] (d)Chatterjee C; McGinty RK; Pellois JP; Muir TW Angew. Chem. Int. Ed. Engl 2007, 46 (16), 2814–8; [PubMed: 17366504] (e)McGinty RK; Kim J; Chatterjee C; Roeder RG; Muir TW Nature 2008, 453 (7196), 812–6. [PubMed: 18449190]

- [52]. Borra R; Camarero JA *Methods Mol. Biol* 2017, 1495, 111–130. [PubMed: 27714613]
- [53]. Charalambous A; Andreou M; Skourides PA *J. Nanobiotechnology* 2009, 7, 9. [PubMed: 20003264]
- [54]. Charalambous A; Andreou M; Antoniadis I; Christodoulou N; Skourides PA *Methods Mol. Biol* 2012, 906, 157–69. [PubMed: 22791430]
- [55]. Dhar T; Kurpiers T; Mootz HD *Methods Mol. Biol* 2011, 131–142.
- [56]. (a)Liu D; Cowburn D *Methods Mol. Biol* 2017, 1495, 131–145; [PubMed: 27714614]
(b)Volkman G; Iwai H *Mol. Biosyst* 2010, 6 (11), 2110–2121. [PubMed: 20820635]
- [57]. Salzmänn M; Pervushin K; Wider G; Senn H; Wuthrich K *Proc. Natl. Acad. Sci. U. S. A* 1998, 95 (23), 13585–90. [PubMed: 9811843]
- [58]. Romanelli A; Shekhtman A; Cowburn D; Muir TW *Proc. Natl. Acad. Sci. U. S. A* 2004, 101 (17), 6397–402. [PubMed: 15087498]
- [59]. Vitali F; Henning A; Oberstrass FC; Hargous Y; Auweter SD; Erat M; Allain FH *EMBO J.* 2006, 25 (1), 150–62. [PubMed: 16362043]
- [60]. (a)Züger S; Iwai H *Nat. Biotechnol* 2005, 23 (6), 736; [PubMed: 15908942] (b)Muona M; Aranko AS; Iwai H *ChemBioChem* 2008, 9 (18), 2958–2961. [PubMed: 19031436]
- [61]. (a)Gupta S; Tycko RJ *Biomol. NMR* 2018, 70 (2), 103–114;(b)Frederick KK; Michaelis VK; Caporini MA; Andreas LB; Debelouchina GT; Griffin RG; Lindquist S *Proc. Natl. Acad. Sci. U. S. A* 2017, 114 (14), 3642–3647. [PubMed: 28330994]
- [62]. (a)Yamazaki T; Otomo T; Oda N; Kyogoku Y; Uegaki K; Ito N; Ishino Y; Nakamura HJ *Am. Chem. Soc* 1998, 120, 5591–5592;(b)Otomo T; Teruya K; Uegaki K; Yamazaki T; Kyogoku YJ *Biomol. NMR* 1999, 14 (2), 105–114;(c)Yagi H; Tsujimoto T; Yamazaki T; Yoshida M; Akutsu HJ *Am. Chem. Soc* 2004, 126 (50), 16632–8.
- [63]. Otomo T; Ito N; Kyogoku Y; Yamazaki T *Biochemistry* 1999, 38 (49), 16040–16044. [PubMed: 10587426]
- [64]. Schubeis T; Nagaraj M; Ritter C *Methods Mol. Biol* 2017, 1495, 147–160. [PubMed: 27714615]
- [65]. Tatulian SA; Qin S; Pande AH; He XJ *Mol. Biol* 2005, 351 (5), 939–47.
- [66]. (a)Jagadish K; Camarero JA *Methods Mol. Biol* 2017, 1495, 41–55; [PubMed: 27714609] (b)Li Y; Bi T; Camarero JA *Adv. Bot. Res* 2015, 76, 271–303; [PubMed: 27064329] (c)Borra R; Camarero JA *Biopolymers* 2013, 100 (5), 502–9. [PubMed: 23893781]
- [67]. (a)Camarero JA; Fushman D; Cowburn D; Muir TW *Bioorg. Med. Chem* 2001, 9 (9), 2479–84; [PubMed: 11553489] (b)Camarero JA; Muir TW *J. Am. Chem. Soc* 1999, 121, 5597–5598.
- [68]. (a)Camarero JA; Fushman D; Sato S; Giriat I; Cowburn D; Raleigh DP; Muir TW *J. Mol. Biol* 2001, 308 (5), 1045–62; [PubMed: 11352590] (b)Schumann FH; Varadan R; Tayakuniyl PP; Grossman JH; Camarero JA; Fushman D *Front. Chem* 2015, 3, 26. [PubMed: 25905098]
- [69]. (a)Iwai H; Lingel A; Pluckthun AJ *Biol. Chem* 2001, 276 (19), 16548–54;(b)Iwai H; Pluckthun A *FEBS Lett.* 1999, (459), 166–172.
- [70]. (a)Gould A; Ji Y; L Aboye T; A Camarero J *Curr. Pharm. Des* 2011, 17 (38), 4294–4307; [PubMed: 22204428] (b)E Garcia A; A Camarero J *Curr. Mol. Pharmacol* 2010, 3 (3), 153–163; [PubMed: 20858197] (c)Jagadish K; Camarero JA *Biopolymers* 2010, 94 (5), 611–6. [PubMed: 20564025]
- [71]. Austin J; Wang W; Puttamadappa S; Shekhtman A; Camarero JA *ChemBioChem* 2009, 10 (16), 2663–70. [PubMed: 19780078]
- [72]. (a)Gould A; Camarero JA *ChemBioChem* 2017;(b)Craik DJ; Du J *Curr. Opin. Chem. Biol* 2017, 38, 8–16. [PubMed: 28249194]
- [73]. Austin J; Kimura RH; Woo YH; Camarero JA *Amino Acids* 2010, 38 (5), 1313–22. [PubMed: 19685144]
- [74]. Gould A; Li Y; Majumder S; Garcia AE; Carlsson P; Shekhtman A; Camarero JA *Mol. Biosyst* 2012.
- [75]. Conibear AC; Wang CK; Bi T; Rosengren KJ; Camarero JA; Craik DJ *J. Phys. Chem. B* 2014.
- [76]. Tavassoli A *Curr. Opin. Chem. Biol* 2017, 38, 30–35. [PubMed: 28258013]
- [77]. Tavassoli A; Benkovic SJ *Nat. Protoc* 2007, 2 (5), 1126–33. [PubMed: 17546003]

- [78]. (a) Jagadish K; Gould A; Borra R; Majumder S; Mushtaq Z; Shekhtman A; Camarero JA *Angew. Chem. Int. Ed. Engl* 2015, 54 (29), 8390–4; [PubMed: 26096948] (b) Jagadish K; Borra R; Lacey V; Majumder S; Shekhtman A; Wang L; Camarero JA *Angew. Chem. Int. Ed. Engl* 2013, 52 (11), 3126–31. [PubMed: 23322720]
- [79]. Deschuyteneer G; Garcia S; Michiels B; Baudoux B; Degand H; Morsomme P; Soumillion P *ACS Chem. Biol* 2010, 5 (7), 691–700. [PubMed: 20527881]
- [80]. (a) Li Y; Aboye T; Breindel L; Shekhtman A; Camarero JA *Biopolymers* 2016, 106 (6), 818–824; [PubMed: 27178003] (b) Bi T; Li Y; Shekhtman A; Camarero JA *Bioorg. Med. Chem* 2018, 26 (6), 1212–1219. [PubMed: 28927803]
- [81]. (a) Ozawa T; Kaihara A; Sato M; Tachihara K; Umezawa Y *Anal. Chem* 2001, 73 (11), 2516–2521; [PubMed: 11403293] (b) Kaihara A; Kawai Y; Sato M; Ozawa T; Umezawa Y *Anal. Chem* 2003, 75 (16), 4176–81; [PubMed: 14632132] (c) Kanno A; Ozawa T; Umezawa Y *Anal. Chem* 2006, 78 (2), 556–60; [PubMed: 16408939] (d) Paulmurugan R; Umezawa Y; Gambhir S *Proc. Natl. Acad. Sci. USA* 2002, 99 (24), 15608–15613. [PubMed: 12438689]
- [82]. (a) Kim SB; Ozawa T; Watanabe S; Umezawa Y *Proc. Natl. Acad. Sci. USA* 2004, 101 (32), 11542–11547; [PubMed: 15289615] (b) Kim SB; Ozawa T; Umezawa Y *Anal. Biochem* 2005, 347 (2), 213–220. [PubMed: 16269125]
- [83]. (a) Ozawa T; Nishitani K; Sako Y; Umezawa Y *Nucleic Acids Res.* 2005, 33 (4), e34–e34; [PubMed: 15731327] (b) Ozawa T; Umezawa Y *Methods Mol. Biol* 2007, 269–280.
- [84]. Kanno A; Ozawa T; Umezawa Y *Anal. Chem* 2006, 78 (23), 8076–8081. [PubMed: 17134142]
- [85]. (a) Skretas G; Wood DW *Protein Sci.* 2005, 14 (2), 523–532; [PubMed: 15632292] (b) Skretas G; Wood DW *J. Mol. Biol* 2005, 349 (3), 464–474; [PubMed: 15878176] (c) Skretas G; Meligova AK; Villalonga-Barber C; Mitsiou DJ; Alexis MN; Micha-Screttas M; Steele BR; Screttas CG; Wood DW *J. Am. Chem. Soc* 2007, 129 (27), 8443–8457; [PubMed: 17569534] (d) Buskirk AR; Ong Y-C; Gartner ZJ; Liu DR *Proc. Natl. Acad. Sci. U. S. A* 2004, 101 (29), 10505–10510. [PubMed: 15247421]
- [86]. Callahan BP; Stanger M; Belfort M *Biotechnol. Bioeng* 2013, 110 (6), 1565–1573. [PubMed: 23280506]
- [87]. Kanno A; Yamanaka Y; Hirano H; Umezawa Y; Ozawa T *Angew. Chem. Int. Ed* 2007, 46 (40), 7595–7599.
- [88]. Mootz HD *ChemBioChem* 2009, 10 (16), 2579–2589. [PubMed: 19708049]
- [89]. Ozawa T; Nogami S; Sato M; Ohya Y; Umezawa Y *Anal. Chem* 2000, 72 (21), 5151–7. [PubMed: 11080857]
- [90]. Huang X; Narayanaswamy R; Fenn K; Szpakowski S; Sasaki C; Costa J; Blancafort P; Lizardi PM *DNA Cell Biol.* 2012, 31 (S1), S-2–S-10. [PubMed: 22313050]
- [91]. Kim SB; Ozawa T; Umezawa Y *Anal. Chem* 2005, 77 (20), 6588–6593. [PubMed: 16223244]
- [92]. (a) Zhang Y; Yang W; Chen L; Shi Y; Li G; Zhou N *Anal. Biochem* 2011, 417 (1), 65–72; [PubMed: 21726524] (b) Lu B; Chen L; Zhang Y; Shi Y; Zhou N *Methods Cell Biol.* 2016, 132, 293–318. [PubMed: 26928549]
- [93]. Zhou J; Wang D; Xi Y; Zhu X; Yang Y; Lv M; Luo C; Chen J; Ye X; Fang L *Biochem. Biophys. Res. Commun* 2017, 488 (4), 621–627. [PubMed: 28501618]
- [94]. Gramespacher JA; Stevens AJ; Nguyen DP; Chin JW; Muir TW *J. Am. Chem. Soc* 2017.
- [95]. Gierach I; Li J; Wu W-Y; Grover GJ; Wood DW *FEBS Open Bio.* 2012, 2, 247–253.
- [96]. Liang R; Zhou J; Liu J *Appl. Environ. Microbiol* 2011, 77 (7), 2488–2495. [PubMed: 21317264]
- [97]. (a) Chin HG; Kim GD; Marin I; Mersha F; Evans TC Jr.; Chen L; Xu MQ; Pradhan S *Proc. Natl. Acad. Sci. U. S. A* 2003, 100 (8), 4510–5; [PubMed: 12671070] (b) Dun BQ; Wang XJ; Lu W; Zhao ZL; Hou SN; Zhang BM; Li GY; Evans TC Jr.; Xu MQ; Lin M *Appl. Environ. Microbiol* 2007, 73 (24), 7997–8000. [PubMed: 17951442]
- [98]. Yang J; Fox GC Jr.; Henry-Smith TV *Proc. Natl. Acad. Sci. U. S. A* 2003, 100 (6), 3513–8. [PubMed: 12629210]
- [99]. Yang J; Henry-Smith TV; Qi M *Transgenic Res.* 2006, 15 (5), 583–93. [PubMed: 16830226]

- [100]. Shen B; Sun X; Zuo X; Shilling T; Apgar J; Ross M; Bougri O; Samoylov V; Parker M; Hancock E; Lucero H; Gray B; Ekborg NA; Zhang D; Johnson JC; Lazar G; Raab RM *Nat. Biotechnol* 2012, 30 (11), 1131–6. [PubMed: 23086202]
- [101]. Hauptmann V; Weichert N; Menzel M; Knoch D; Paegle N; Scheller J; Spohn U; Conrad U; Gils M *Transgenic Res.* 2013, 22 (2), 369–77. [PubMed: 23001519]
- [102]. (a)Zhu F; Liu Z; Chi X; Qu H *Sci. China Life Sci* 2010, 53 (6), 683–689; [PubMed: 20602271]
(b)Zhu F; Liu Z; Wang X; Miao J; Qu H; Chi X *Sci. China Life Sci* 2013, 56 (3), 262–7. [PubMed: 23526393]
- [103]. Wang P; Chen T; Sakurai K; Han BX; He Z; Feng G; Wang F *Sci. Rep* 2012, 2, 497. [PubMed: 22773946]
- [104]. Truong D-JJ; Kühner K; Kühn R; Werfel S; Engelhardt S; Wurst W; Ortiz O *Nucleic Acids Res.* 2015, 43 (13), 6450–6458. [PubMed: 26082496]
- [105]. Carvajal-Vallejos P; Pallisse R; Mootz HD; Schmidt SR *J. Biol. Chem* 2012, 287 (34), 28686–96. [PubMed: 22753413]
- [106]. Stevens AJ; Sekar G; Shah NH; Mostafavi AZ; Cowburn D; Muir TW *Proc. Natl. Acad. Sci. USA* 2017, 114 (32), 8538–8543. [PubMed: 28739907]
- [107]. Kimura RH; Tran AT; Camarero JA *Angew. Chem. Int. Ed* 2006, 45 (6), 973–6.
- [108]. Camarero JA; Kimura RH; Woo YH; Shekhtman A; Cantor J *ChemBioChem* 2007, 8 (12), 1363–6. [PubMed: 17590879]

Figure 1.

Schematics of intein-mediated protein splicing. **A.** The DNA encoding the protein precursor is transcribed into mRNA, and then translated into the precursor protein. The protein undergoes a self-catalytic modification which results in the excision of the intein and a mature protein that is produced by ligating the N- and C-extein fragments. **B.** Scheme representing intein-mediated protein splicing in *trans*. Protein trans-splicing involves an intein self-processing domain that is split into N- (I_N) and C-intein (I_C) fragments. These fragments cannot splice individually, however, under the appropriate conditions they can bind to each other with high specificity forming a fully active intein domain in *trans*. Protein trans-splicing can be also mediated by naturally-occurring or artificially split inteins.

Figure 2.

Canonical conserved sequence motifs found in inteins that facilitate the protein splicing. The protein splicing domain or Intein is shown as a color blue, and N- and C-exteins exteins are show in green and grey, respectively. The canonical motifs and corresponding blocks are shown on the intein domain. It is also shown the location of the homing endonuclease domain (HED) and the split site on naturally occurring split-inteins and the homing endonuclease domain (HED).

Figure 3.

Scheme showing the canonical protein splicing mechanism for a Cys intein. When the intein contains a Ser or Ser/Thr residues at the position +1 of the intein and C-extein, respectively, the reaction works the formation of the corresponding ester intermediates. The protein splicing reaction involves first a nucleophilic attack of the Cys thiol group of the intein to the peptide bond at the N-extein-intein junction through an N→S acyl shift (step 1). The resulting linear thioester intermediate reacts with the thiol group of Cys residue at position +1 on the C-extein to produce a branched thioester intermediate (step 2). The side-chain of the Asn at the C-terminal position cleaves the peptide bond at the intein-N-extein junction to form an aspartimide and a reactive thioester intermediate (step 3). The thioester intermediate

then rearranges into a more stable peptide bond through an S->N acyl shift to produce the final spliced protein (step 4).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 4.

Scheme showing the intein-mediated protein purification by using an intein modified to favor N-terminal cleavage. The protein of interest (POI) is fused to the N-terminus of engineered intein containing a C-terminal tag sequence, in this case a chitin binding domain (CBD) although other compatible affinity tags may be also used. The corresponding fusion protein is expressed and purified using an appropriate affinity column. Cleavage of the POI from the column is induced by thiolysis using an appropriate thiol-containing compound like for example sodium 2-mercaptoethanesulfonate (MESNA).

Figure 5.

Schematic representations of intein-mediated protein labeling. **A.** Site-specific labeling and fluorescence activation of a protein of interest (POI) by making use of a FRET-quenched protein trans-splicing [28]. The introduction of a fluorescence quencher (Q) into the I_C polypeptide (blue) blocks the fluorescence signal of the fluorophore (F) located at the C-extein (orange). The sequence CFNK stand for the first four residues of the DnaE C-extein, i.e. Cys-Phe-Asn-Lys. When protein trans-splicing occurs, the fluorophore is covalently attached to the C-terminus of the POI triggering its fluorescence. **B.** In-cell segmental isotopic labeling of the N-terminal part of a POI using the NMR-active ¹⁵N isotope [60a]. LB stands for Luria Bertani media which contains natural isotopic distribution of C and N.

Figure 6.

Intein-mediated backbone cyclization of a polypeptide linear precursor. The intein-cyclotide construct is recombinantly expressed and can be cyclized and folded either *in vitro* [74, 107] or in cell [71, 108]. The intein used in this approach is modified to inhibit C-terminal cleavage. This is usually accomplished by mutating the last Asn residue on the block G of the intein to Ala.

Figure 7.

Use of PTS for the production of backbone cyclized polypeptides. This approach when combined with efficient split-inteins can be used for the efficient in-cell production of genetically-encoded libraries of backbone cyclized polypeptides [80b].

A. Protein-protein Interactions **B. Protein-DNA Interactions** **C. Protein localization and Internalization** **D. Protease Activity** **E. Redox state** **F. Allosteric-driven intein-based biosensor** **Figure 8.**

Schematic representations of the different applications of inteins in the design of molecular biosensors. **A.** Use of an intein-based system to sense protein-protein interactions. Binding between two protein activates splicing resulting in an activated reporter. **B.** Intein-mediated biosensor to detect protein-DNA interactions. The binding of the Zn-finger domains (ZF1/ZF2) to their specific target sites induces intein-mediated protein splicing, which results in production of a functional full-length luciferase that generates luminescence in the presence of luciferin. **C.** Protein localization and internalization biosensor. Translocation of a POI into the nucleus is detected when luciferase is generated through PTS. **D.** Intein-mediated biosensor for protease activity. An inactive and backbone-cyclized version of luciferase containing a protease cleavage site is cleaved in the presence of the corresponding

proteolytic enzyme to produce a fully active linear enzyme. **E.** Intein-mediated biosensor of redox state. Under oxidizing conditions, the biosensor is normally activated (FRET on) as the Cys in the block A of the intein (N-terminal Cys) is trapped forming a disulfide bond. Under more reductive conditions the disulfide bond is reduced activating the intein, which results in the cleavage of the cyan fluorescent protein (CFP) from the intein-yellow fluorescence protein (YFP) fusing decreasing the corresponding FRET signal. Single letter coding is used to indicate the Cys-Pro-Gly-Cys loop in the redox regulatory segment of the intein N-terminal region. **F.** Intein-mediated allosteric biosensor for a small molecule. Binding of a small molecule triggers in a conformational change of the intein resulting in its activation and production of the corresponding reporter.