



Review

# Inteins: Localized Distribution, Gene Regulation, and Protein Engineering for Biological Applications

Theetha L. Pavankumar

Department of Microbiology and Molecular Genetics, Briggs Hall, One Shields Ave, University of California, Davis, CA 95616, USA; pavan@ucdavis.edu; Tel.: +1-530-754-9702; Fax: +1-530-754-8973

Received: 12 February 2018; Accepted: 25 February 2018; Published: 28 February 2018

**Abstract:** Inteins are self-splicing polypeptides with an ability to excise themselves from flanking host protein regions with remarkable precision; in the process, they ligate flanked host protein fragments. Inteins are distributed sporadically across all three domains of life (bacteria, archaea, and unicellular eukaryotes). However, their apparent localized distribution in DNA replication, repair, and recombination proteins (the 3Rs), particularly in bacteria and archaea, is enigmatic. Our understanding of the localized distribution of inteins in the 3Rs, and their possible regulatory role in such distribution, is still only partial. Nevertheless, understanding the chemistry of post-translational self-splicing of inteins has opened up opportunities for protein chemists to modify, manipulate, and bioengineer proteins. Protein-splicing technology is adapted to a wide range of applications, starting with untagged protein purification, site-specific protein labeling, protein biotinylation, isotope incorporation, peptide cyclization, as an antimicrobial target, and so on. This review is focused on the chemistry of splicing; the localized distribution of inteins, particularly in the 3Rs and their possible role in regulating host protein function; and finally, the use of protein-splicing technology in various protein engineering applications.

**Keywords:** inteins; splicing; Hint domains; DNA replication; repair and recombination; protein engineering

## 1. Introduction

Inteins are intervening polypeptides with an ability to splice themselves out from the flanking protein fragments (exteins) post-translationally. Inteins (intervening proteins) were first found associated with the *VMA1* gene (also known as *TFP1*) that encodes for  $\alpha$  subunit of vacuolar membrane  $H^+$ -translocating adenosine triphosphatase ( $H^+$ -ATPase) of *Saccharomyces cerevisiae* [1]. The TFP1 gene product was observed to be larger (118.6 kDa) than the estimated 69 kDa, with N- and C-terminal regions being homologous to the similar  $H^+$ -ATPases [1]. Later, work by Kane et al. convincingly demonstrated that the TFP1 precursor protein (118.6 kDa) splices itself out post-translationally into 69-kDa ( $H^+$ -ATPase) and 50-kDa proteins [2]. Since then, several hundreds of inteins have been reported [3].

Inteins are widely dispersed in nature. Their broad phylogenic distribution across all three domains of life suggests that inteins have ancient origins. Despite their wide distribution, they are absent in multicellular organisms. Inteins are somehow analogous to introns. The splicing of inteins can occur either spontaneously or under favorable conditions. In most cases, inteins are expressed as a single contiguous polypeptide (*cis*-splicing inteins); in some instances, they are transcribed and translated separately as split or *trans*-splicing inteins. These *trans*-splicing inteins re-associate via the zipper-like interface and perform a splicing reaction [4–6]. Both contiguous and split intein splicing processes and their chemistry of making and breaking peptide bonds have been exploited to develop various protein engineering techniques.

## 2. Hint Domain Superfamily and Different Forms

Hint (Hedge-Hog/Intein) is a protein domain containing fundamental characteristic features needed for the protein-splicing reaction to occur. Hint domains exist in different forms. They are 140 to 160 amino acid long polypeptides containing 4–6 conserved motifs. They are primarily comprised of three  $\beta$ -sheets and two  $\alpha$ -helices linked through loop regions. Hint domain was shown to fold into a horseshoe-like core structure with a pseudo-two-fold symmetry [7]. The proper folding of the Hint protein domain is all that seems to be an essential criterion for the precise protein splicing process to happen, and thus, Hint domains are the key players in the protein-splicing process. The Hint domain superfamily is comprised of three important subfamilies: (1) Intein-Hint domain, (2) Hedgehog-Hint domain, and (3) Bacterial intein-like (BIL)-Hint domains. These subfamilies are largely varied by the type of functional domains associated with them and the way they splice-out.

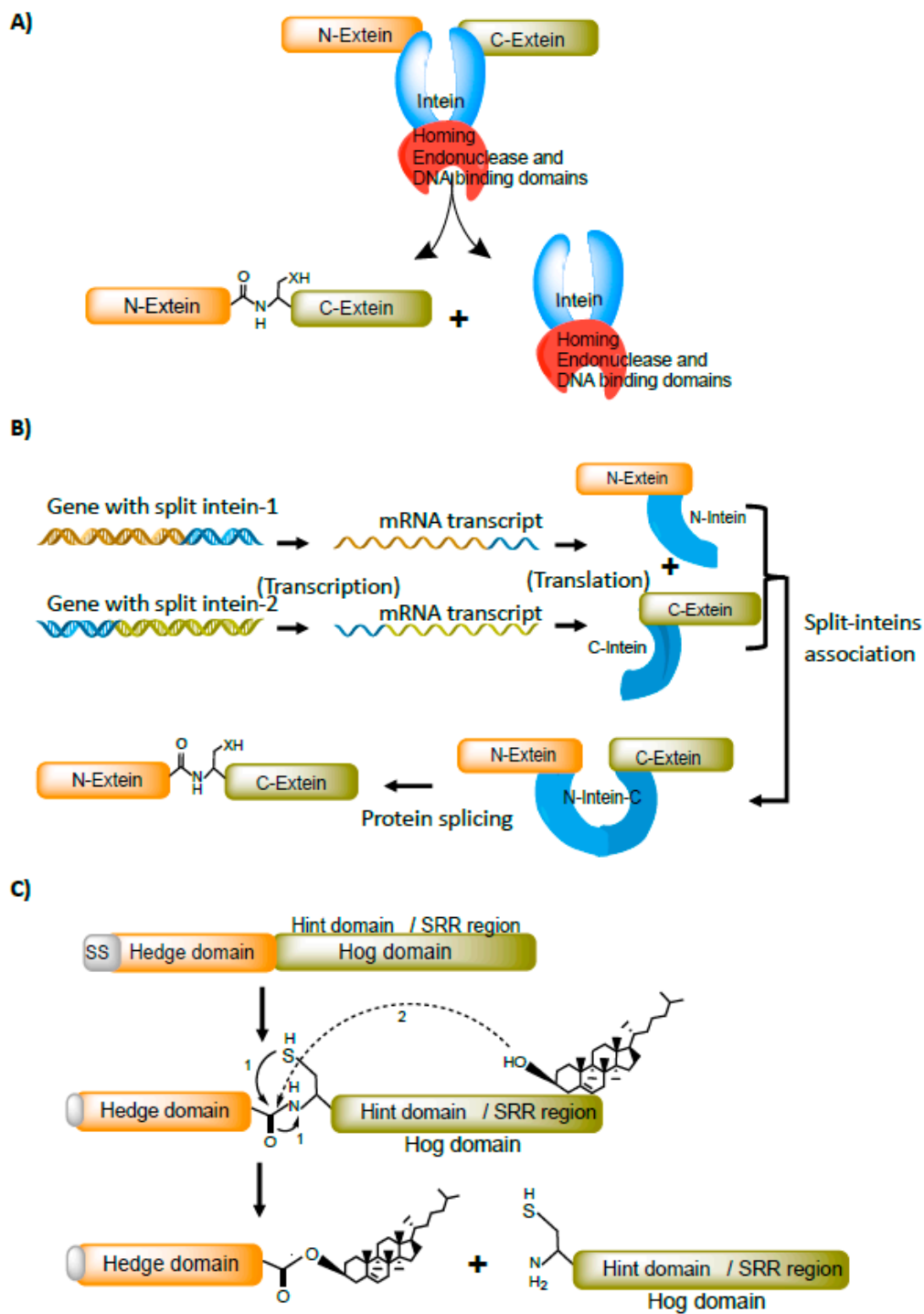
## 3. Intein-Hint Domain

Inteins are part of the Hint domain superfamily. All inteins have a protein-splicing Hint domain. However, inteins vary widely by having different functional domains embedded within them. Many inteins have homing endonuclease (HED) and DNA binding domains embedded within them. Inteins with HED domain splice out spontaneously from the precursor protein by ligating N-terminus extein (N-extein) with C-terminus extein (C-extein) to form a functional protein (Figure 1A). Inteins embedded with HED and DNA binding domains are selfish genetic mobile elements. They recognize a nucleotide sequence ranging from 14–40 bp and induce sequence-specific double-strand breaks (DSBs) on an intein/HED free allele [8]. The intein/HED domain is then mobilized into an intein/HED free allele by DSB-mediated DNA repair via homologous recombination. However, many inteins have apparently lost the HED and DNA binding domains during evolution. Inteins with or without HED and DNA binding domains are contiguous and are produced in a single polypeptide form.

Unlike contiguous inteins, certain inteins are produced in two or more polypeptide parts. These split inteins are transcribed and translated separately. Although split inteins are being produced separately, they retain all the hint domain features needed for the splicing process [4,9,10]. They associate in-trans and perform a splicing reaction (Figure 1B). Split inteins are naturally found in cyanobacteria [9,11]. Intriguingly, these are inserted in a conserved motif of essential genes, particularly, in DNA replication machinery [9]. It is speculated that split inteins resulted from genomic arrangements that split the contiguous intein into two functionally independent entities during evolution. The proper Hint domain folding is the essential parameter for splicing to occur in both contiguous and split inteins.

## 4. Hedgehog-Hint Domain

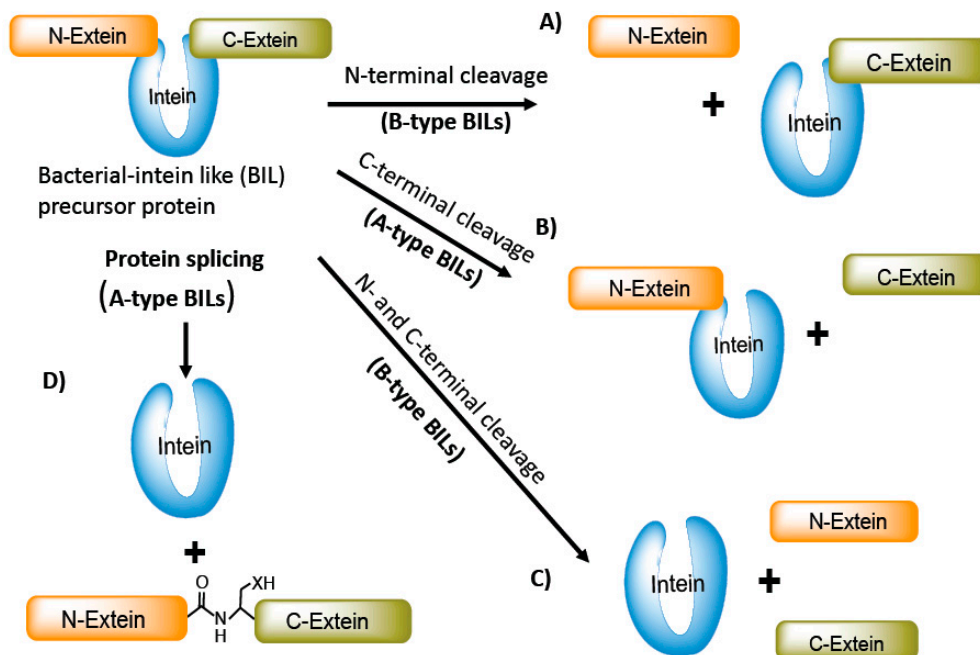
Hedgehog signaling proteins are the key regulators of the developmental-signaling pathway in metazoans. Hedgehog proteins are composed of three domains. The N-terminal region of the protein is the Hedge domain and, the C-terminal region (Hog) is comprised of a Hint domain and a sterol recognition region (SRR). The C-terminal region with Hint domain has the same structural fold as inteins [7] and is responsible for the maturation of the N-terminus Hedge protein. During the maturation process, the conserved glycine and cysteine residues of Hog domain are rearranged to form a thioester. The hydroxyl-oxygen of the cholesterol molecule attached to the SRR region attacks the thioester bond. During the attack, the cholesterol molecule gets transferred on to the Hedge protein via ester linkage and the modified N-terminus region (Hedge domain), with cholesterol at C-terminus splice-out from the C-terminal Hog-hint region [12] (Figure 1C).



**Figure 1.** Different forms of the hint domain. (A) Inteins containing homing endonuclease (HE) and DNA binding domains. Inteins with HED are selfish genetic mobile elements capable of integrating into the alleles. (B) Split inteins are *trans*-splicing elements that are transcribed and translated separately and have the ability to associate *in-trans*, and splice out. (C) Hedgehog proteins with hint domain (hog domain). The conserved glycine and cysteine residues rearrange to form a thioester bond (step-1; N to S acyl rearrangement). The hydroxyl group of cholesterol attacks thioester bond (step-2; *trans*-esterification) and links cholesterol to the hedge domain releasing hog domain. SS, a processed signal-sequence peptide at the N-terminus of the Hedge domain.

## 5. Bacterial-Intein-Like Hint Domain

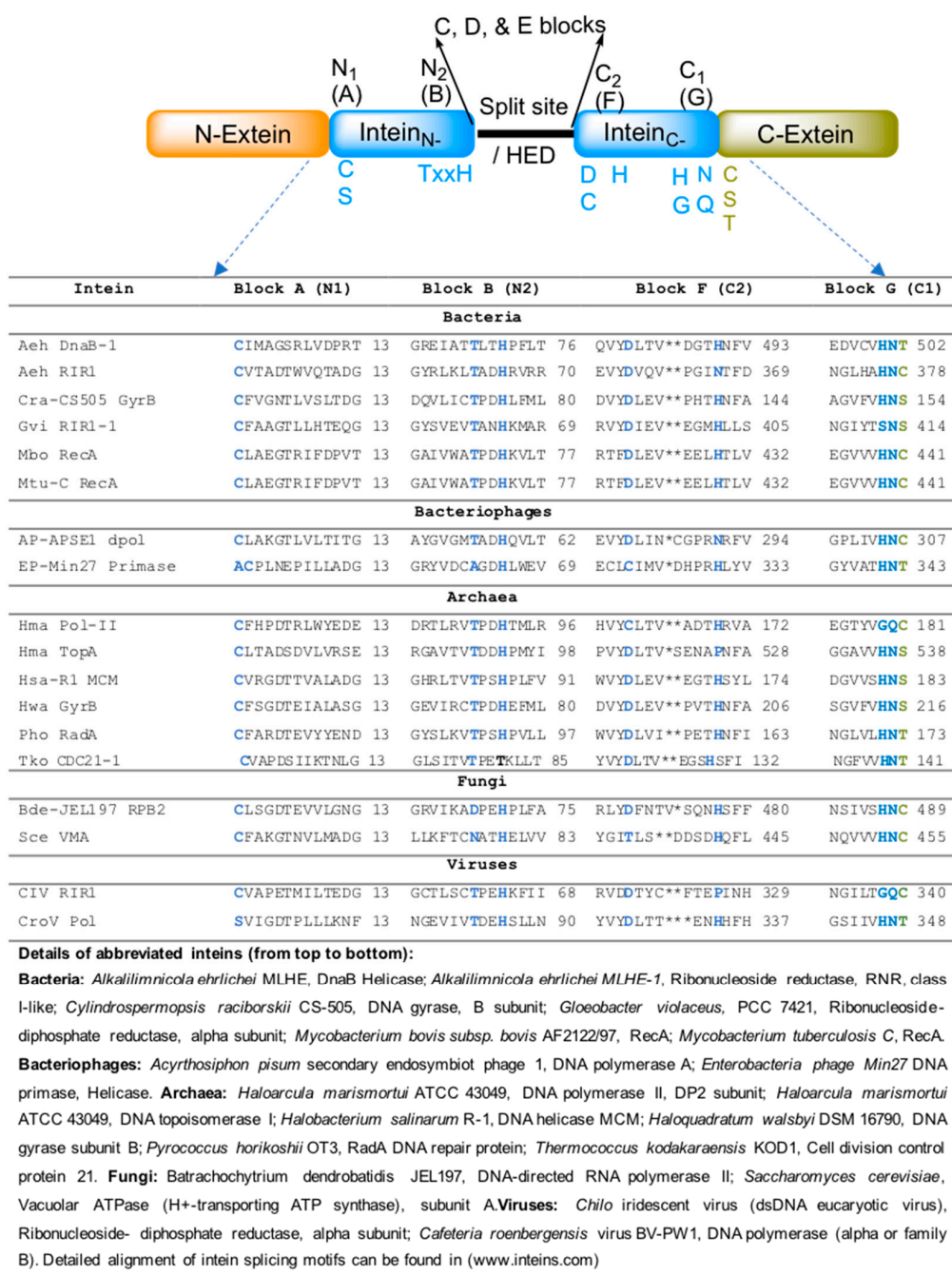
Bacterial-intein-like (BIL) domains are similar to intein and Hedgehog-hint domain family proteins but differ in certain sequence features [13]. The phylogenetic distribution and type of host protein in which the BILs are inserted also vary compared to the intein and Hog-Hint domain family. Unlike inteins, BIL domains are inserted in non-conserved variable regions of proteins of divergent bacteria. Two different types of BIL domains have been identified: A- and B-type. Both A- and B-type domains can self-cleave at the N- or C-terminus of the BIL domains. The A-type BIL domains, in some instances, follow an alternative splicing mechanism, whereas B-type domains appear to follow the canonical protein-splicing mechanism [14] (Figure 2).



**Figure 2.** Bacterial intein-like (BIL) proteins and their splicing behavior. BILs precursor proteins follow the non-canonical splicing mechanism. A-type BILs often lack nucleophilic residue at the C-terminal flanking region and are susceptible to C-terminal cleavage. B-type BILs undergo both C-terminal and/or N-terminal cleavage (A,C). A-type can splice completely to produce fully functional protein and often undergo C-terminal cleavage (B,D).

## 6. Mechanism of Intein-Mediated Protein Splicing

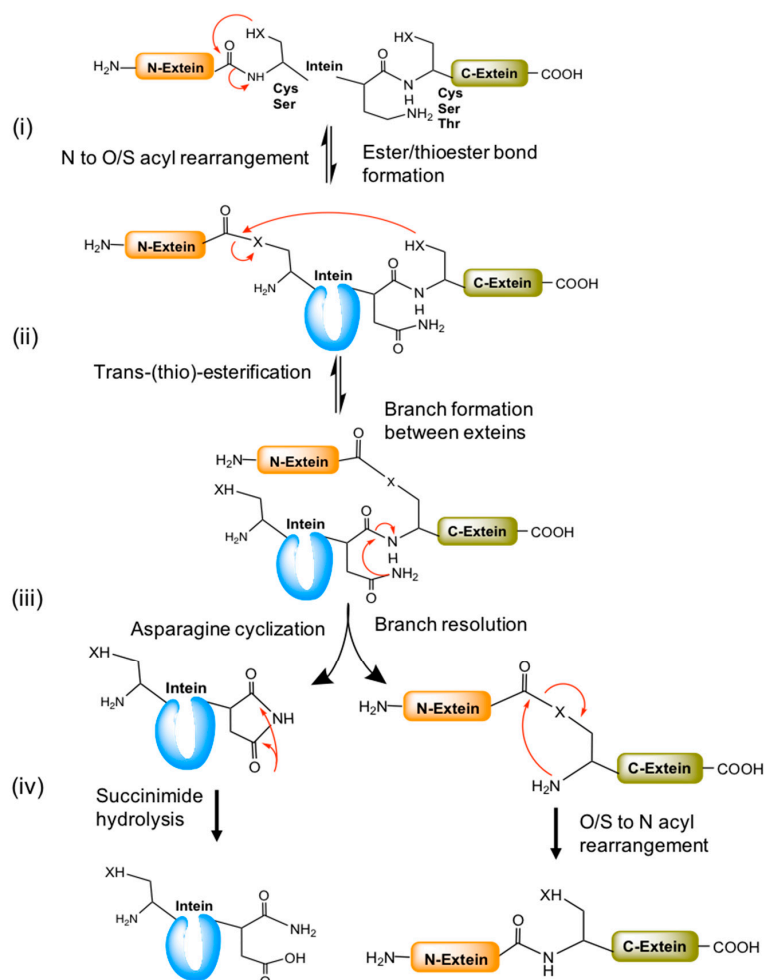
Protein splicing is all about breaking a peptide bond (at intein–extein junctions) and forming a new peptide bond between extein polypeptides, but in a meticulous way. It is a precise chemical reaction that occurs between specific conserved amino acids residing within inteins and exteins. Protein splicing is typically a single-turnover reaction and, in general, does not require cofactors. Sequence alignment of inteins obtained from intein database ([www.inteins.com](http://www.inteins.com)) revealed that inteins have four important splicing motifs, each comprising highly conserved similar amino acids. These motifs are further designated into seven blocks: A to G. Among the seven blocks, A, B, F, and G (also known as N1, N2, C2, and C1, respectively) are the essential blocks and reside within the intein region (Figure 3).



**Figure 3.** Important motifs and conserved splicing elements of inteins. Inteins contain four important motifs: The N-terminal region with N<sub>1</sub> and N<sub>2</sub>; and the C-terminal region with C<sub>1</sub> and C<sub>2</sub> motifs. The important splicing amino acid residues of each motif are indicated. The conserved amino acid residues (C, S, and T) on the C-extein are also indicated. These amino acids are also important for the splicing process to occur. A list of selected intein-associated proteins of bacteria, bacteriophages, viruses, and fungi, and their protein sequence alignment of conserved A, B, E, and F motifs, is also given.

Splicing mechanisms of different classes of inteins have been investigated in detail. However, it is hard to define a universal mechanism by which the protein-splicing reaction works. In some instance, it detours from canonical splicing pathway as some conserved functionally similar amino acids need assistance from the adjacent amino acids to complete the splicing reaction. A generalized mechanism of canonical splicing involves the following steps;

1. Ester/thioester bond formation by N to O/S acyl rearrangement,
2. Trans-(thio)-esterification between exteins,
3. Asparagine cyclization, and
4. Spontaneous hydrolysis of the amino-succinamide residue and O/S to N acyl rearrangement to form a peptide bond between exteins (Figure 4).



**Figure 4.** A general mechanism of the canonical splicing pathway. (i) Splicing begins with ester/thio-esterification (serine or cysteine residues) by N to O/S acyl rearrangement at the N-extein end. (ii) The esterified N-extein is linked to the C-extein by *trans*-(thio) esterification (by C, S or T). The branch resolution and Asn cyclization occurs in the next step with the help of conserved penultimate histidine residues. (iii) Later, the N- and C-exteins undergo O/S to N acyl rearrangement to form a new peptide bond between them. (iv) The cyclized asparagine converted into succinimide by hydrolysis. Red arrows represent the chemical reaction/s between the functional groups indicated.

The splicing process can occasionally deviate from canonical splicing as a consequence of variation in the conserved amino acids of the N1, N2, C2, and C1 blocks (Blocks A, B, F, and G, respectively) (Figure 3). The lack of block-N2 or penultimate histidine in block-C1 affects the splicing process. For example, the *Thermococcus kodakaraensis* CDC21-1 (TkoCDC-21-1) intein has threonine instead of histidine at block-N2 [15]. In this case, a lysine residue (K58) residing outside of the standard intein conserved motifs seems to catalyze the initial N-S acylation reaction [15]. Similarly, the DnaE inteins of cyanobacteria have either serine or alanine in place of histidine at block-C1 [9]. These histidine residues are important for initiating N to O/S acyl rearrangement and asparagine

cyclization, respectively [15,16]. In addition, an external nucleophilic attack (such as water and thiol) on the (thio) ester intermediate can cause N-terminal cleavage. Similarly, C-terminal cleavage could also occur if asparagine cyclization occurs during the delay or absence of transesterification.

## 7. Localized Distribution of Inteins and Post-Translational Regulation

Inteins are distributed sporadically across all three domains of life. About 24%, 47%, and 1% of the total genome of bacteria, archaea, and lower eukaryotes, respectively, contain inteins [17]. Intriguingly, 62% and 67% of inteins of bacteria and archaea reside in DNA replication, recombination, and repair proteins. In particular, inteins are embedded in DNA polymerases, DNA topoisomerases, DNA helicases, DNA strand exchange proteins, and ribonucleotide reductases. Inteins are often found inserted at critical functional sites such as the phosphate-binding loop (P-loop) or the catalytic or ligand binding sites of many essential DNA replication and repair proteins [18]. Also, the insertion site is varied across the host proteins. DNA helicases (such as DnaB, PcrA, and UvrD) and DNA strand-exchange proteins (such as RecA and RadA) have inteins inserted at the P-loop, and DNA polymerases and topoisomerases have found intein insertion at catalytic or ligand binding sites [18–20]. Inteins are also found in the ribonucleotide reductase large subunit of the *Chilo* iridescent virus [21]. A list of proteins with intein insertion is given in Table 1.

**Table 1.** Some of the intein-containing proteins involved in the DNA replication, recombination, and repair process.

Proteins	Function	Category	Organism
RecA/RadA	DNA-strand exchange	DNA repair	Eubacteria/Archaea
DnaB	Replicative DNA helicase	DNA replication	Eubacteria
Rad54/SWI-SNF2	dsDNA dependent ATPase	SF2 helicases & Chromatin remodeler	Eubacteria
UvrD/Rep/PcrA	ATP dependent DNA helicase (SF1 family)	DNA mismatch repair	Eubacteria
DnaE, Catalytic $\alpha$ -subunit of DNA pol III	DNA polymerase	DNA Replication	Eubacteria
Bacterial DNA polymerase I (PolA)	DNA synthesis	DNA Replication	Eubacteria
Bacterial DNA polymerase II (PolB)	DNA synthesis	DNA Replication	Eubacteria
DNA polymerase III $\tau$ and $\gamma$	DNA synthesis	DNA Replication	Eubacteria
RuvB	APT dependent DNA helicase	Holliday junction branch migration	Eubacteria
Ribonucleoside diphosphate reductase	Ribonucleotide diphosphate reductase	DNA synthesis	Eubacteria; Archaea
DNA gyrase subunit A & B	Stabilizing the DNA (DNA replication, resection)	Topoisomerase	Eubacteria
DnaG	DNA primase	DNA replication	Eubacteria
RecG	ATP dependent DNA helicase	DNA replication	Eubacteria
Replication factor-C small unit (RFC)	DNA clamp loader	DNA Replication	Archaea
DNA polymerase II large unit (PolC/DP2)	DNA synthesis	DNA Replication	Archaea
DNA polymerase II small unit (PolB)	DNA synthesis	DNA Replication	Archaea
Mini-chromosome maintenance protein (MCM)	Replicative DNA helicase	DNA replication	Archaea

The discovery of enigmatic distribution and preferred insertion site of inteins has left scientists with more questions than answers.

1. Why do inteins localize in DNA replication, recombination, and repair proteins, and in RNA transcription machinery?
2. Is there any significance of intein association with particular conserved regions (like P-loop of NTPases or catalytic sites) of essential proteins?
3. Do they have a regulatory role in DNA replication machinery under stress conditions? or
4. Is it just a selective pressure that retains intein association with particular regions of certain functional proteins, as argued earlier [17]?

Still, the localized distribution and preferred insertion of intein into essential genes is still a matter of debate. However, recent advances in intein biology indicate that inteins may act as

environmental cues. The pathogenic bacteria such as *Mycobacterium tuberculosis* and *Mycobacterium leprae* have intein inserted in *recA* gene [22]. Inteins inserted into *M. tuberculosis* and *M. leprae* are different in size, sequence, and location of insertion. Importantly, the *in-vivo* splicing of both inteins also varies. *M. tuberculosis* RecA precursor protein splice out spontaneously when expressed in *E. coli* and also, when N- and C-terminal purified fragments are provided *in-trans* in the presence of DTT [23]. In contrast, *M. leprae* RecA precursor protein could splice out only in native cells and in *M. smegmatis*, not in *E. coli* [22,24]. This observation suggests that the splicing of *M. leprae* RecA precursor protein is a host-specific reaction and requires an unknown splicing cofactor specific to *M. leprae* and *M. smegmatis*.

There is growing evidence of the existence of conditional protein splicing, in which the splicing process is regulated by many factors such as redox state, temperature, pH, and DNA. It is observed that some inteins contain two cysteines at the active splicing site and have the potential to form a disulfide bridge. In the case of the *Pyrococcus abyssi* DNA polymerase II intein precursor, these two cysteine residues form an intramolecular disulfide bond that inhibits protein splicing [25]. It is speculated that *P. abyssi* is anaerobic and the presence of oxygen may pose oxidative stress. The oxidative stress may promote disulfide bridge formation and thus inhibit DNA polymerase II precursor splicing during oxidative stress. Similarly, the SufB intein of *M. tuberculosis* is found to act as a sensor for oxidative and nitrosative stress [26].

Protein splicing is also regulated by temperature. An intein inserted at the ATPase domain of RadA (a DNA strand-exchange protein) of the hyperthermophilic archaeon *Pyrococcus horikoshii* seems to be regulated in a temperature-dependent manner. It is demonstrated that efficient splicing occurs only at temperatures between 65 to 85 °C and is controlled by native exteins (the host RadA protein fragments) [27]. *P. horikoshii*, being a hyperthermophilic archaeon, temperature-dependent modulation of RadA production, may serve as a regulatory mechanism in this archaeon.

Deoxyribonucleic acids (DNA) are the essential genetic elements of living organisms. A recent study indicates that ssDNA (single-stranded DNA) and DNA damage response both seem to trigger the protein splicing of the *P. horikoshii* RadA protein. Interestingly, the stimulation is specific to ssDNA but insensitive to dsDNA (double-stranded DNA), dNTPs, and RNA [28]. All the evidence given above points to the post-translational regulation of a protein's function through splicing. Therefore, it can be envisaged that protein splicing is a post-translational regulatory mechanism, by which biological functions are regulated to allow the individual to thrive under various stress conditions.

## 8. Applications of Intein Splicing in Protein Engineering and Biological Applications

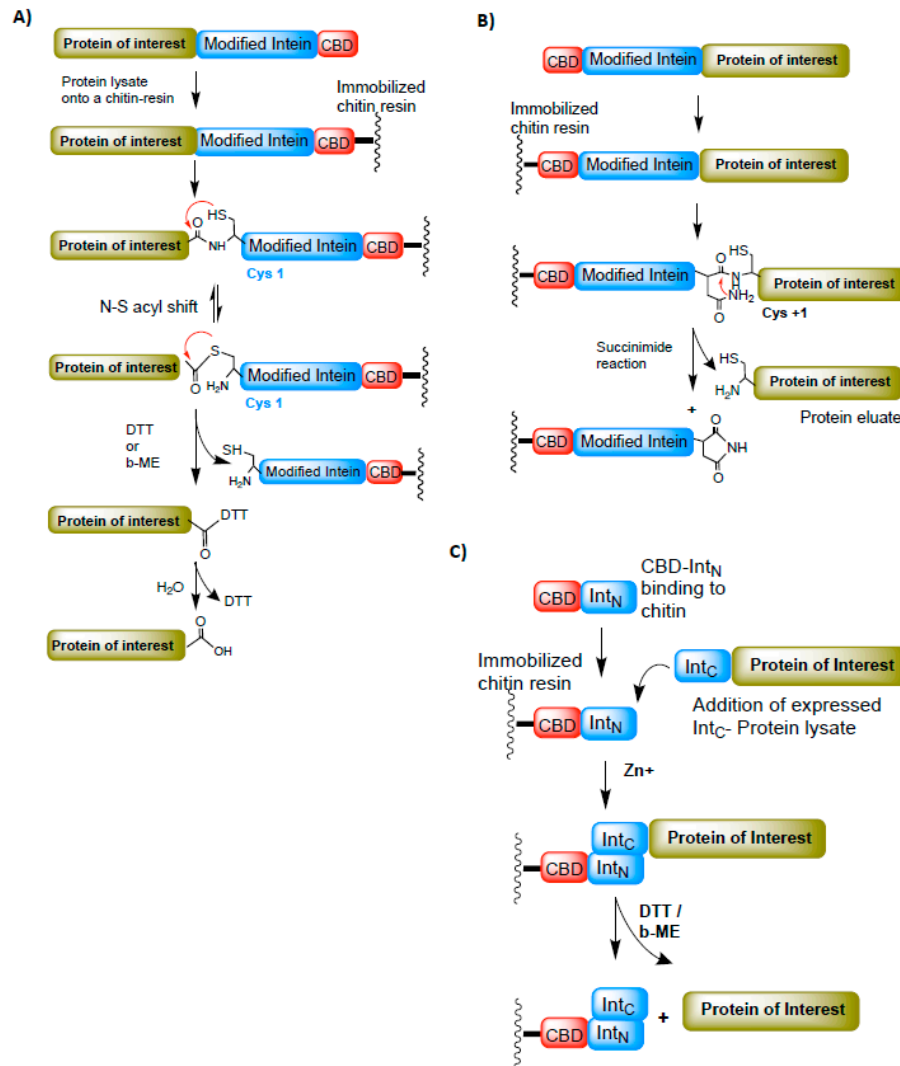
The intein-mediated way of breaking and making a peptide bond offers promising scope for protein engineering. Since the discovery of inteins and their splicing chemistry, protein chemists have been striving to adapt and develop a wide variety of bioengineering techniques for biological applications. To date, the splicing technique has been successfully employed in protein purification, protein modifications, peptide cyclization, and as intein-based biosensor and reporter systems. Inteins are also being investigated as a potential target for antimicrobial drugs.

## 9. Inteins in Protein Purification

Affinity chromatography is a widely adopted technique for protein purification. It requires affinity tags such as 6XHis, FLAG octapeptide, Glutathione S-transferase (GST), Maltose-binding protein (MBP), or Chitin-binding protein (CBP) to be attached either at the N- or C-terminus of the protein of interest. Often, these tags interfere with a protein's function and need to be removed by proteases (such as precision protease, Thrombin, or Factor Xa). In some instances, affinity-tag removal by proteases is time-consuming and may leave additional amino acids on the protein of interest. The discovery of intein-mediated protein splicing process enabled the purification of recombinant proteins without affinity tags (fused affinity tags are cleaved from the protein of interest during intein-mediated splicing) and with no additional amino acids on them, as demonstrated by Chang et al. in 1997 [29]. In this study, a modified intein from *Saccharomyces cerevisiae* (*Sce* VMA intein) was fused



to the chitin-binding domain (CBD) from *Bacillus circulans* as an affinity tag. In general, the protein of interest is cloned at the N-terminus of the intein-CBD fusion domain and immobilized on the chitin column (Figure 5A). The splicing (cleavage) is induced by adding either DTT or  $\beta$ -ME. This technique was further extended to fuse the N-terminus of the protein of interest with the C-terminus intein and purify the proteins by the C-terminus splicing reaction [30] (Figure 5B). However, in this case, cysteine (Cys<sup>+1</sup>) on the protein of interest (Extein) is required for the splicing process to occur efficiently.



**Figure 5.** The intein-based protein purification techniques. (A) A method to purify the protein of interest (POI). The C-terminus of POI is fused to the N-terminus intein-CBD (chitin-binding domain). The POI-Intein-CBD fusion is expressed and immobilized onto the chitin resin. The POI is later cleaved from the bound intein-CBD-chitin resin by adding DTT or  $\beta$ -ME. (B) A modified intein-based purification method to purify N-terminus-fused POI attached to the C-terminus of intein-CBD domain. In this case, the POI is cleaved from the fused domain by succinimide hydrolysis. (C) Split intein-based purification system. The N-terminus intein (int<sub>N</sub>) fused to CBD is immobilized onto the chitin resin and the C-terminus intein (int<sub>C</sub>) fused to POI is allowed to trans-associate with int<sub>C</sub>-CBD domain. Zinc is added to minimize the splicing reaction during the binding process. Later, POI is cleaved off from the bound fractions by adding DTT or  $\beta$ -ME. Red arrows represent the chemical reaction/s between the functional groups indicated.

Similar to the Sce-VMA intein-mediated purification system, a *Mycobacterium xenopi* Gyrase-A intein containing a purification technique was also developed. Here temperature is used as a splicing factor to cleave an immobilized target protein [31]. Nevertheless, preventing or minimizing the cleavage of protein precursor during expression and rapid cleavage of protein of interest during purification are the essential parameters for a successful intein-mediated purification system. Since then, many investigators have tested and developed efficient purification systems using modified inteins fused to different affinity tags [32–36].

The split intein chemistry also contributed remarkably to the protein purification system. Their ability to associate *in-trans* and splice out rapidly have added a next level to the purification of recombinant proteins. Many engineered split inteins are employed to develop better and more efficient purification systems [4,10,37–42]. One such purification system developed using an engineered split intein from *Nostoc punctiforme* DnaE is found to be efficient and robust in cleaving the protein of interest at a remarkably faster rate [4,43,44]. The steps involved in protein purification using *N. punctiforme* DnaE split inteins are illustrated in Figure 5C.

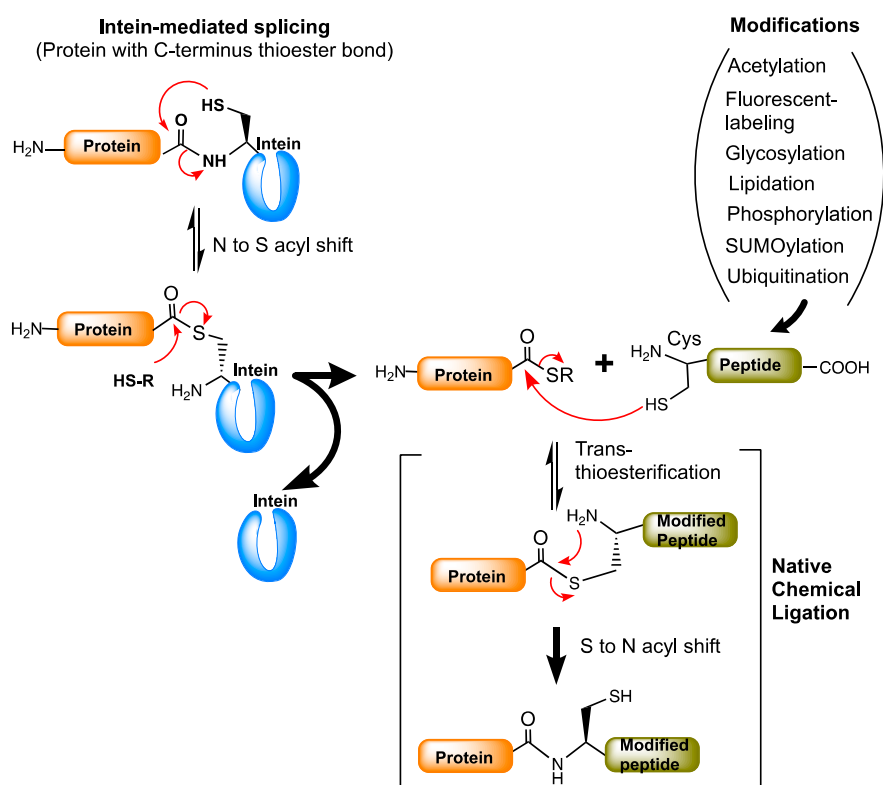
## 10. Protein Modifications Using Splicing Chemistry

The chemistry of ligating two peptides goes back to the 1990s. A Native Chemical Ligation (NCL) method developed by Kent and co-workers has been found to be tremendously useful in peptide chemistry [45]. It is a reaction that ligates two peptides by allowing a reaction between a peptide with C-terminal thioester ( $\alpha$ -thioester) and a peptide with a cysteine at the N-terminus end. NCL is a technically challenging method and is limited to producing polypeptides up to 15 kDa. The limitations of NCL were eliminated by intein-mediated protein splicing chemistry-based methods such as expressed protein ligation (EPL) and protein trans-splicing (PTS). Expressed protein ligation [46,47] is essentially similar to NCL but involves an intein-based splicing reaction to produce the protein of interest with a C-terminal thioester to which cysteine-containing chemically modified proteins or peptides are ligated under mild aqueous conditions (Figure 6). The protein trans-synthesis [48] is very similar to the split inteins splicing process (Figure 1B), but in the case of PTS the N- and C-exteins are the proteins of interest with modifications. Both EPL and PTS are used for linking two proteins of interest with desired modifications for biological applications. Both techniques have proven useful in post-translational modifications (PTMs) of proteins.

Expressed protein ligation is the most common method used for the semisynthesis of post-translationally modified proteins. EPL has been used for protein modifications such as acetylation, phosphorylation, ubiquitination, SUMOylation, glycosylation, lipidation, and in the fluorescent labeling of proteins [49]. EPL is widely used in the phosphorylation of protein kinase Csk [47] to study the structural basis of heteromeric Smad protein [50,51], in the preparation of glycoproteins [52], for ubiquitination of PCNA (proliferating cell nuclear antigen) [53], and in histone proteins [54]. It is also employed in site-specific labeling of proteins for the Förster resonance energy transfer (FRET) studies [55]. More detailed biological applications of intein-based EPL and PTS methods are reviewed in [49,56].

Cyclized proteins or peptides are found naturally in bacteria, plants, and mammals. Cyclization is a process of linking the N-terminus of a peptide (or protein) with its C-terminus end by a peptide bond. Cyclosporine is one such cyclized peptide found in fungi and is being used as an immunosuppressant in the medical field. Cyclization of proteins or peptide seems to offer stability, improved biological activity, and affinity compared to their counterparts. Many naturally existing cyclotides (cyclized peptides) have antimicrobial, anticancer, and anthelmintic properties. The cyclization of a target protein can be carried out by sandwiching the target protein between the N- and C-inteins, as shown by using a split intein of the *dnaE* gene from *Synechocystis* species PCC6803 [57]. Polypeptide cyclization is shown to produce biologically active, fast-folding, and denaturation-resistant recombinant proteins compared to their counterparts [58]. Hence, cyclization of synthetic peptides is a subject of interest in pharmaceutical industries. Peptide cyclization of c-Crk,  $\beta$ -lactamase, and green fluorescent protein

(GFP) was performed to obtain better derivatives [59–61]. Detailed information on protein cyclization and applications of both EPL and PTS in peptide cyclization is given in [62].



**Figure 6.** Protein modifications by expressed protein ligation (EPL). The protein of interest fused to the intein is expressed and purified in esterified form. The esterified protein is then attached to the modified peptides (or proteins) by trans-(thio)esterification process using native chemical ligation (NCL). Red arrows represent the chemical reaction/s between the functional groups indicated.

Intein- and split-intein-mediated splicing chemistry is further extended to in vivo applications such as developing an intein-based biosensor (redox-state sensor in *E. coli*), in sensing DNA methylation, to understand site-specific proteolysis and study protein–protein interactions [49,56,63].

## 11. Inteins as Microbial Drug Target

Many inteins reside in pathogenic bacteria, such as *Mycobacterium tuberculosis*, *Mycobacterium leprae*, and *Coxiella burnetii*. Inteins are also found in pathogenic fungi like *C. neoformans*, *C. gattii*, and *Histoplasma capsulatum*. As inteins are largely found in essential genes of many organisms including *M. tuberculosis* and *M. leprae*, inhibiting intein splicing may offer a target for antimicrobial drugs. Cisplatin, an anticancer drug, has recently been shown to inhibit the protein-splicing process in mycobacteria [64]. Nevertheless, cisplatin cannot be used as an antimicrobial agent as it is being used as an anticancer drug and may pose severe side effects if used as an antimicrobial agent. At the least, intein splicing inhibition offers a novel microbial drug target, particularly for the treatment of tuberculosis and leprosy.

## 12. Conclusions

The localized distribution of inteins, particularly in DNA replication and repair proteins at the critical sites, is intriguing. Twenty-five years of intein research has provided plentiful information on the nature of inteins, their distribution, mechanism of splicing, and use in biological applications. However, scientists must still look for possible intein-mediated regulation of essential genes and their

role in microbial adaptation to stress conditions. Although recent studies have revealed a possible intein-mediated regulation of genes and their probable role in adaptation, further research is needed to understand and interpret their existence and distribution. The intein-mediated splicing mechanism (of making and breaking peptide bonds) has provided a tremendous tool to adapt in biological applications and also technological development. Nevertheless, there is no limit to the usability of the splicing mechanism in future applications.

**Acknowledgments:** I thank Anurag Sinha, Manjunatha Belaganahalli, Naofumi Handa, and Satish Kallappagoudar for their critical reading and comments on the manuscript.

**Conflicts of Interest:** The author declares no conflict of interest.

## Abbreviations

HED	Homing-endonuclease domain
3Rs	replication, repair and recombination
BIL	bacterial intein-like
SRR	sterol recognition region
DTT	dithiothreitol
NTPase	nucleotide tri-phosphatases
$\beta$ -ME	beta-Mercaptoethanol
NCL	native chemical ligation
EPL	expressed protein ligation
PTS	protein trans-synthesis
CDB	chitin-binding domain

## References

- Hirata, R.; Ohsumk, Y.; Nakano, A.; Kawasaki, H.; Suzuki, K.; Anraku, Y. Molecular structure of a gene, *VMA1*, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **1990**, *265*, 6726–6733. [[PubMed](#)]
- Kane, P.; Yamashiro, C.; Wolczyk, D.; Neff, N.; Goebel, M.; Stevens, T. Protein splicing converts the yeast *TFPI* gene product to the 69-kd subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* **1990**, *250*, 651–657. [[CrossRef](#)] [[PubMed](#)]
- Perler, F.B. Inbase: The intein database. *Nucleic Acids Res.* **2002**, *30*, 383–384. [[CrossRef](#)] [[PubMed](#)]
- Zettler, J.; Schutz, V.; Mootz, H.D. The naturally split *Npu DnaE* intein exhibits an extraordinarily high rate in the protein trans-splicing reaction. *FEBS Lett.* **2009**, *583*, 909–914. [[CrossRef](#)] [[PubMed](#)]
- Liu, X.Q.; Yang, J. Split *dnaE* genes encoding multiple novel inteins in *Trichodesmium erythraeum*. *J. Biol. Chem.* **2003**, *278*, 26315–26318. [[CrossRef](#)] [[PubMed](#)]
- Sorci, M.; Dassa, B.; Liu, H.; Anand, G.; Dutta, A.K.; Pietrokovski, S.; Belfort, M.; Belfort, G. Oriented covalent immobilization of antibodies for measurement of intermolecular binding forces between zipper-like contact surfaces of split inteins. *Anal. Chem.* **2013**, *85*, 6080–6088. [[CrossRef](#)] [[PubMed](#)]
- Hall, T.M.; Porter, J.A.; Young, K.E.; Koonin, E.V.; Beachy, P.A.; Leahy, D.J. Crystal structure of a hedgehog autoprocessing domain: Homology between hedgehog and self-splicing proteins. *Cell* **1997**, *91*, 85–97. [[CrossRef](#)]
- Chevalier, B.S.; Stoddard, B.L. Homing endonucleases: Structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res.* **2001**, *29*, 3757–3774. [[CrossRef](#)] [[PubMed](#)]
- Caspi, J.; Amitai, G.; Belenkiy, O.; Pietrokovski, S. Distribution of split *dnaE* inteins in cyanobacteria. *Mol. Microbiol.* **2003**, *50*, 1569–1577. [[CrossRef](#)] [[PubMed](#)]
- Iwai, H.; Zuger, S.; Jin, J.; Tam, P.H. Highly efficient protein trans-splicing by a naturally split *dnaE* intein from *Nostoc punctiforme*. *FEBS Lett.* **2006**, *580*, 1853–1858. [[CrossRef](#)] [[PubMed](#)]
- Wu, H.; Hu, Z.; Liu, X.Q. Protein trans-splicing by a split intein encoded in a split *dnaE* gene of *Synechocystis* sp. PCC6803. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 9226–9231. [[CrossRef](#)] [[PubMed](#)]
- Burglin, T.R. The hedgehog protein family. *Genome Biol.* **2008**, *9*, 241. [[CrossRef](#)] [[PubMed](#)]

13. Amitai, G.; Belenkiy, O.; Dassa, B.; Shainskaya, A.; Pietrokovski, S. Distribution and function of new bacterial intein-like protein domains. *Mol. Microbiol.* **2003**, *47*, 61–73. [[CrossRef](#)] [[PubMed](#)]
14. Dassa, B.; Haviv, H.; Amitai, G.; Pietrokovski, S. Protein splicing and auto-cleavage of bacterial intein-like domains lacking a C'-flanking nucleophilic residue. *J. Biol. Chem.* **2004**, *279*, 32001–32007. [[CrossRef](#)] [[PubMed](#)]
15. Tori, K.; Cheriyan, M.; Pedomallu, C.S.; Contreras, M.A.; Perler, F.B. The *Thermococcus kodakaraensis* Tko CDC21-1 intein activates its N-terminal splice junction in the absence of a conserved histidine by a compensatory mechanism. *Biochemistry* **2012**, *51*, 2496–2505. [[CrossRef](#)] [[PubMed](#)]
16. Frutos, S.; Goger, M.; Giovani, B.; Cowburn, D.; Muir, T.W. Branched intermediate formation stimulates peptide bond cleavage in protein splicing. *Nat. Chem. Biol.* **2010**, *6*, 527–533. [[CrossRef](#)] [[PubMed](#)]
17. Novikova, O.; Jayachandran, P.; Kelley, D.S.; Morton, Z.; Merwin, S.; Topilina, N.I.; Belfort, M. Intein clustering suggests functional importance in different domains of life. *Mol. Biol. Evol.* **2016**, *33*, 783–799. [[CrossRef](#)] [[PubMed](#)]
18. Novikova, O.; Topilina, N.; Belfort, M. Enigmatic distribution, evolution, and function of inteins. *J. Biol. Chem.* **2014**, *289*, 14490–14497. [[CrossRef](#)] [[PubMed](#)]
19. Goodwin, T.J.; Butler, M.I.; Poulter, R.T. Multiple, non-allelic, intein-coding sequences in eukaryotic RNA polymerase genes. *BMC Biol.* **2006**, *4*, 38. [[CrossRef](#)] [[PubMed](#)]
20. Swithers, K.S.; Senejani, A.G.; Fournier, G.P.; Gogarten, J.P. Conservation of intron and intein insertion sites: Implications for life histories of parasitic genetic elements. *BMC Evol. Biol.* **2009**, *9*, 303. [[CrossRef](#)] [[PubMed](#)]
21. Pietrokovski, S. Identification of a virus intein and a possible variation in the protein-splicing reaction. *Curr. Biol.* **1998**, *8*, R634–R635. [[CrossRef](#)]
22. Davis, E.O.; Thangaraj, H.S.; Brooks, P.C.; Colston, M.J. Evidence of selection for protein introns in the recAs of pathogenic *Mycobacteria*. *EMBO J.* **1994**, *13*, 699–703. [[PubMed](#)]
23. Mills, K.V.; Lew, B.M.; Jiang, S.; Paulus, H. Protein splicing in trans by purified N- and C-terminal fragments of the *Mycobacterium tuberculosis recA* intein. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 3543–3548. [[CrossRef](#)] [[PubMed](#)]
24. Frischkorn, K.; Springer, B.; Bottger, E.C.; Davis, E.O.; Colston, M.J.; Sander, P. In vivo splicing and functional characterization of *Mycobacterium leprae recA*. *J. Bacteriol.* **2000**, *182*, 3590–3592. [[CrossRef](#)] [[PubMed](#)]
25. Chen, W.; Li, L.; Du, Z.; Liu, J.; Reitter, J.N.; Mills, K.V.; Linhardt, R.J.; Wang, C. Intramolecular disulfide bond between catalytic cysteines in an intein precursor. *J. Am. Chem. Soc.* **2012**, *134*, 2500–2503. [[CrossRef](#)] [[PubMed](#)]
26. Topilina, N.I.; Green, C.M.; Jayachandran, P.; Kelley, D.S.; Stanger, M.J.; Piazza, C.L.; Nayak, S.; Belfort, M. *Sufb* intein of mycobacterium tuberculosis as a sensor for oxidative and nitrosative stresses. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 10348–10353. [[CrossRef](#)] [[PubMed](#)]
27. Topilina, N.I.; Novikova, O.; Stanger, M.; Banavali, N.K.; Belfort, M. Post-translational environmental switch of RadA activity by extein-intein interactions in protein splicing. *Nucleic Acids Res.* **2015**, *43*, 6631–6648. [[CrossRef](#)] [[PubMed](#)]
28. Lennon, C.W.; Stanger, M.; Belfort, M. Protein splicing of a recombinase intein induced by ssDNA and DNA damage. *Genes Dev.* **2016**, *30*, 2663–2668. [[CrossRef](#)] [[PubMed](#)]
29. Chong, S.; Mersha, F.B.; Comb, D.G.; Scott, M.E.; Landry, D.; Vence, L.M.; Perler, F.B.; Benner, J.; Kucera, R.B.; Hirvonen, C.A.; et al. Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. *Gene* **1997**, *192*, 271–281. [[CrossRef](#)]
30. Chong, S.; Montello, G.E.; Zhang, A.; Cantor, E.J.; Liao, W.; Xu, M.Q.; Benner, J. Utilizing the c-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step. *Nucleic Acids Res.* **1998**, *26*, 5109–5115. [[CrossRef](#)] [[PubMed](#)]
31. Southworth, M.W.; Amaya, K.; Evans, T.C.; Xu, M.Q.; Perler, F.B. Purification of proteins fused to either the amino or carboxy terminus of the *Mycobacterium xenopi* gyrase a intein. *Biotechniques* **1999**, *27*, 110–114. [[PubMed](#)]
32. Banki, M.R.; Feng, L.; Wood, D.W. Simple bioseparations using self-cleaving elastin-like polypeptide tags. *Nat. Methods* **2005**, *2*, 659–661. [[CrossRef](#)] [[PubMed](#)]
33. Banki, M.R.; Gerngross, T.U.; Wood, D.W. Novel and economical purification of recombinant proteins: Intein-mediated protein purification using in vivo polyhydroxybutyrate (PHB) matrix association. *Protein Sci.* **2005**, *14*, 1387–1395. [[CrossRef](#)] [[PubMed](#)]

34. Wang, Z.; Li, N.; Wang, Y.; Wu, Y.; Mu, T.; Zheng, Y.; Huang, L.; Fang, X. Ubiquitin-intein and SUMO2-intein fusion systems for enhanced protein production and purification. *Protein Expr. Purif.* **2012**, *82*, 174–178. [[CrossRef](#)] [[PubMed](#)]
35. Wood, D.W.; Derbyshire, V.; Wu, W.; Chartrain, M.; Belfort, M.; Belfort, G. Optimized single-step affinity purification with a self-cleaving intein applied to human acidic fibroblast growth factor. *Biotechnol. Prog.* **2000**, *16*, 1055–1063. [[CrossRef](#)] [[PubMed](#)]
36. Xie, Y.G.; Luan, C.; Zhang, H.W.; Han, F.F.; Feng, J.; Choi, Y.J.; Groleau, D.; Wang, Y.Z. Effects of thioredoxin: SUMO and intein on soluble fusion expression of an antimicrobial peptide OG2 in *Escherichia coli*. *Protein Pept. Lett.* **2013**, *20*, 54–60. [[CrossRef](#)] [[PubMed](#)]
37. Aranko, A.S.; Oeemig, J.S.; Zhou, D.; Kajander, T.; Wlodawer, A.; Iwai, H. Structure-based engineering and comparison of novel split inteins for protein ligation. *Mol. Biosyst.* **2014**, *10*, 1023–1034. [[CrossRef](#)] [[PubMed](#)]
38. Lu, W.; Sun, Z.; Tang, Y.; Chen, J.; Tang, F.; Zhang, J.; Liu, J.N. Split intein facilitated tag affinity purification for recombinant proteins with controllable tag removal by inducible auto-cleavage. *J. Chromatogr. A* **2011**, *1218*, 2553–2560. [[CrossRef](#)] [[PubMed](#)]
39. Shah, N.H.; Dann, G.P.; Vila-Perello, M.; Liu, Z.; Muir, T.W. Ultrafast protein splicing is common among cyanobacterial split inteins: Implications for protein engineering. *J. Am. Chem. Soc.* **2012**, *134*, 11338–11341. [[CrossRef](#)] [[PubMed](#)]
40. Stevens, A.J.; Brown, Z.Z.; Shah, N.H.; Sekar, G.; Cowburn, D.; Muir, T.W. Design of a split intein with exceptional protein splicing activity. *J. Am. Chem. Soc.* **2016**, *138*, 2162–2165. [[CrossRef](#)] [[PubMed](#)]
41. Stevens, A.J.; Sekar, G.; Shah, N.H.; Mostafavi, A.Z.; Cowburn, D.; Muir, T.W. A promiscuous split intein with expanded protein engineering applications. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 8538–8543. [[CrossRef](#)] [[PubMed](#)]
42. Wong, S.; Mosabbir, A.A.; Truong, K. An engineered split intein for photoactivated protein trans-splicing. *PLoS ONE* **2015**, *10*, e0135965. [[CrossRef](#)] [[PubMed](#)]
43. Ramirez, M.; Valdes, N.; Guan, D.; Chen, Z. Engineering split intein *dnaE* from nostoc punctiforme for rapid protein purification. *Protein Eng. Des. Sel.* **2013**, *26*, 215–223. [[CrossRef](#)] [[PubMed](#)]
44. Guan, D.; Chen, Z. Affinity purification of proteins in tag-free form: Split intein-mediated ultrarapid purification (sirp). *Methods Mol. Biol.* **2017**, *1495*, 1–12. [[PubMed](#)]
45. Dawson, P.E.; Muir, T.W.; Clark-Lewis, I.; Kent, S.B. Synthesis of proteins by native chemical ligation. *Science* **1994**, *266*, 776–779. [[CrossRef](#)] [[PubMed](#)]
46. Evans, T.C., Jr.; Benner, J.; Xu, M.Q. Semisynthesis of cytotoxic proteins using a modified protein splicing element. *Protein Sci.* **1998**, *7*, 2256–2264. [[CrossRef](#)] [[PubMed](#)]
47. Muir, T.W.; Sondhi, D.; Cole, P.A. Expressed protein ligation: A general method for protein engineering. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6705–6710. [[CrossRef](#)] [[PubMed](#)]
48. Paulus, H. Protein splicing and related forms of protein autoprocessing. *Ann. Rev. Biochem.* **2000**, *69*, 447–496. [[CrossRef](#)] [[PubMed](#)]
49. Vila-Perello, M.; Muir, T.W. Biological applications of protein splicing. *Cell* **2010**, *143*, 191–200. [[CrossRef](#)] [[PubMed](#)]
50. Chacko, B.M.; Qin, B.Y.; Tiwari, A.; Shi, G.; Lam, S.; Hayward, L.J.; De Caestecker, M.; Lin, K. Structural basis of heteromeric Smad protein assembly in TGF-beta signaling. *Mol. Cell* **2004**, *15*, 813–823. [[CrossRef](#)] [[PubMed](#)]
51. Wu, J.W.; Hu, M.; Chai, J.; Seoane, J.; Huse, M.; Li, C.; Rigotti, D.J.; Kyin, S.; Muir, T.W.; Fairman, R.; et al. Crystal structure of a phosphorylated SMAD2. Recognition of phosphoserine by the MH2 domain and insights on SMAD function in TGF-beta signaling. *Mol. Cell* **2001**, *8*, 1277–1289. [[CrossRef](#)]
52. Buskas, T.; Ingale, S.; Boons, G.J. Glycopeptides as versatile tools for glycobiology. *Glycobiology* **2006**, *16*, 113R–136R. [[CrossRef](#)] [[PubMed](#)]
53. Chen, J.; Ai, Y.; Wang, J.; Haracska, L.; Zhuang, Z. Chemically ubiquitylated PCNA as a probe for eukaryotic translesion DNA synthesis. *Nat. Chem. Biol.* **2010**, *6*, 270–272. [[CrossRef](#)] [[PubMed](#)]
54. Chatterjee, C.; Muir, T.W. Chemical approaches for studying histone modifications. *J. Biol. Chem.* **2010**, *285*, 11045–11050. [[CrossRef](#)] [[PubMed](#)]
55. Maag, D.; Fekete, C.A.; Gryczynski, Z.; Lorsch, J.R. A conformational change in the eukaryotic translation preinitiation complex and release of EIF1 signal recognition of the start codon. *Mol. Cell* **2005**, *17*, 265–275. [[CrossRef](#)] [[PubMed](#)]

56. Wood, D.W.; Camarero, J.A. Intein applications: From protein purification and labeling to metabolic control methods. *J. Biol. Chem.* **2014**, *289*, 14512–14519. [[CrossRef](#)] [[PubMed](#)]
57. Evans, T.C., Jr.; Martin, D.; Kolly, R.; Panne, D.; Sun, L.; Ghosh, I.; Chen, L.; Benner, J.; Liu, X.Q.; Xu, M.Q. 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. [[CrossRef](#)] [[PubMed](#)]
58. Camarero, J.A.; Fushman, D.; Sato, S.; Giriat, I.; Cowburn, D.; Raleigh, D.P.; Muir, T.W. Rescuing a destabilized protein fold through backbone cyclization. *J. Mol. Biol.* **2001**, *308*, 1045–1062. [[CrossRef](#)] [[PubMed](#)]
59. Camarero, J.A.; Muir, T.W. Biosynthesis of a head-to-tail cyclized protein with improved biological activity. *J. Am. Chem. Soc.* **1999**, *121*, 5597–5598. [[CrossRef](#)]
60. Iwai, H.; Lingel, A.; Pluckthun, A. Cyclic green fluorescent protein produced in vivo using an artificially split PI-*Pfui* intein from *Pyrococcus furiosus*. *J. Biol. Chem.* **2001**, *276*, 16548–16554. [[CrossRef](#)] [[PubMed](#)]
61. Iwai, H.; Pluckthun, A. Circular beta-lactamase: Stability enhancement by cyclizing the backbone. *FEBS Lett.* **1999**, *459*, 166–172. [[CrossRef](#)]
62. Aboye, T.L.; Camarero, J.A. Biological synthesis of circular polypeptides. *J. Biol. Chem.* **2012**, *287*, 27026–27032. [[CrossRef](#)] [[PubMed](#)]
63. Topilina, N.I.; Mills, K.V. Recent advances in in vivo applications of intein-mediated protein splicing. *Mob. DNA* **2014**, *5*. [[CrossRef](#)] [[PubMed](#)]
64. Zhang, L.; Zheng, Y.; Callahan, B.; Belfort, M.; Liu, Y. Cisplatin inhibits protein splicing, suggesting inteins as therapeutic targets in *Mycobacteria*. *J. Biol. Chem.* **2011**, *286*, 1277–1282. [[CrossRef](#)] [[PubMed](#)]



© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).