# Mechanistic and Kinetic Considerations of Protein Splicing

Thomas C. Evans, Jr. and Ming-Qun Xu*

*New England Biolabs, Inc., 32 Tozer Road, Beverly, Massachusetts 01915*

*Received March 22, 2002*

## Contents



## I. Introduction

Protein splicing is a post-translational processing event in which an internal protein segment (termed an intein) is removed from a precursor protein with the concomitant ligation of the flanking protein sequences[1] (termed exteins). Only the intein and the amino acid residue immediately downstream of the C-terminal splice junction are essential for self-splicing activity. Furthermore, the experimentally characterized inteins required no exogenous cofactors or energy sources such as ATP or GTP. This does not preclude the existence of protein splicing elements that rely on cofactors. A number of inteins do not splice in vitro or in a foreign host organism or protein. This may be due to folding difficulties or, of greater interest, may indicate that they are not self-sufficient. Introns, which are removed at the RNA level by self-splicing or by a spliceosome complex,[2] have many superficial similarities to inteins. Perhaps, there is a spliceosome-like assembly for intein removal from a precursor protein.

* To whom correspondence should be addressed. Phone: (978) 927-5054. Fax: (978) 921-1350. E-mail: xum@neb.com.

During the protein splicing reaction peptide bonds at the intein N- and C-termini are broken to release it from the precursor protein. Bond breakage is mediated by formation of (thio)ester intermediates and facilitated by a putative oxyanion hole and/or a potential catalytic triad, such as certain proteases. Unlike proteases, inteins also ligate two flanking protein segments through a native peptide bond as part of their normal reaction scheme. As is often the case for natural systems, this activity elegantly couples bond breakage and formation as part of the splicing mechanism.

Since their original discovery, over 100 inteins have been listed in InBase, the intein registry.[3] Most were identified by searching for unexplained gene insertions or known protein splicing motifs. It is worth noting that only ca. 30% of the known inteins have been tested for protein splicing activity. Inteins are present in all three domains of life, but none have been discovered in the nuclear genome of a multicellular organism. However, they are present in chloroplasts and viral or bacterial pathogens of multicellular organisms. With one possible exception,[4] inteins do not appear to benefit the host and are considered to be "selfish genes". Intein evolution will not be covered in this review, but this topic has been eloquently discussed in previous papers.[5−9] The current review will focus on the types, biochemistry, and application of inteins.

## II. Intein History and Classification

### A. Discovery

In 1990 two groups reported the existence of protein splicing elements. Serendipitously, they both discovered the intervening sequence present in a *Saccharomyces cerevisiae* vacuolar $H^+$-ATPase subunit.[10,11] This protein splicing element was later termed the *Sce* VMA1 intein.[1] The discovery occurred during the cloning and characterization of a new $H^+$-ATPase. Unexpectedly, the sequenced gene predicted a 119 kDa protein instead of the 70 kDa molecular mass of the purified subunit. The N- and C-termini of the predicted amino acid sequence showed over 70% identity to vacuolar $H^+$-ATPases from other organisms, while the middle portion displayed no significant homology. This central region was found to have approximately 30% sequence identity to the yeast HO endonuclease.

From the accumulated evidence it appeared that two proteins, an $H^+$-ATPase and an HO endonu-

10.1021/cr9601369 CCC: $39.75 
Published on Web 11/15/2002

Thomas C. Evans, Jr., received his Ph.D. in biochemistry from the University of Minnesota in 1996. His graduate work on the characterization of calcium-dependent membrane binding proteins was carried out under the mentorship of Dr. Gary Nelsestuen. He then joined the laboratory of Dr. Ming-Qun Xu at New England Biolabs, Inc., as a postdoctoral associate. There he developed intein-based methodologies for protein engineering. In 1999 Tom was hired as a Staff Scientist at New England Biolabs, Inc., and in 2002 was promoted to Senior Scientist. His laboratory is currently involved in the kinetic analysis of protein splicing, applying inteins in the prevention of transgene spread, and investigating novel protein cyclization pathways.

Ming-Qun Xu received his B.Sc. in cell biology and genetics at the University of Sciences and Technology of China. He received his Ph.D. in 1989 under the supervision of Professor David A. Shub at the State University of New York at Albany. His doctoral work was on a new self-splicing group I intron in bacteriophage T4. From 1989 to 1992, he worked as a postdoctoral research associate with Professor David Shub on the evolution and molecular mechanisms of RNA self-splicing. He joined Dr. Francine Perler in 1992 as a postdoctoral research associate and studied protein splicing elements (inteins) from thermophilic archaeal organisms at New England Biolabs, Inc. In 1994, Dr. Xu joined New England Biolabs as a Staff Scientist and investigated the mechanism of protein splicing and its applications in protein purification and ligation. He was promoted to Senior Scientist in 1997. His current interests include the structure and function of protein splicing elements, engineering of self-splicing inteins for protein expression and purification, applying inteins in ligation, labeling, and cyclization of polypeptides in vivo and in vitro, and antigen preparation and antibody purification by intein-mediated protein ligation.

clease, were encoded in the same gene. Splicing at the RNA level would explain the removal of an intervening nucleic acid sequence. Two factors indicated that RNA splicing was not the explanation: (i) the detectable mRNA was in the full, unspliced form, and (ii) the highly conserved sequences found in yeast introns were absent.

Further experimental evidence indicated that splicing was not at the RNA level, but instead at the protein level. A stop codon introduced into the intervening DNA sequence by deleting two base pairs resulted in a truncated $H^+$-ATPase. Another single base pair deletion that restored the reading frame prior to the stop codon produced the mature, full-length $H^+$-ATPase. RNA splicing would not be sensitive to the reading frame. Also, protein expression in a rabbit reticulocyte lysate system resulted in the generation of a 69 kDa $H^+$-ATPase and a second 50 kDa protein, presumably representing the excised intervening protein segment.

Shortly after the first protein splicing publication a number of new potential protein splicing elements were reported.[12,13] Sequence comparison between these proteins permitted the assignment of conserved residues flanking the N- and C-terminal splice junctions[14,15] (Figure 1). These conserved residues were required for proper protein splicing. In addition, splicing motifs were identified[14−16] that facilitated intein discovery. Interestingly, no known intein is encoded in the nuclear genome of a multicellular organism. One has been found in the ribonucleotide reductase large subunit of an insect virus, *Chilo* iridescent virus.[17] The majority of newly discovered inteins are similar to the first protein splicing elements. However, two new types, the so-called alanine inteins[18] and the naturally occurring *trans*-splicing intein,[4] show a variation on a theme.

## B. Intein Nomenclature/Types

Initially the *S. cerevisiae* protein splicing element was simply referred to as a "spacer".[11] Later, new splicing elements were termed intervening protein sequences. The discovery of multiple protein splicing elements made a flexible and consistent nomenclature necessary. In response, a collection of researchers studying these unique proteins outlined a naming system.[1] It was modeled after RNA splicing so that the process was termed protein splicing and the unspliced chimera was referred to as the precursor protein. The spacer was renamed an intein from *in*ternal pro*tein* sequence. Following the same convention, the amino acid sequences flanking the intein were termed *ex*ternal pro*tein* sequences or exteins. The extein fused to the intein N- or C-terminus is the N- or C-extein, respectively. A system of naming inteins from different organisms and different genes was also defined. Inteins are named using the genus and species designations (typically the three letter genus/species abbreviation) followed by the gene name. For example, the protein splicing element found in the *S. cerevisiae* membrane vacuolar $H^+$-ATPase was termed the *Sce* VMA1 intein.

Comparison of biochemical studies of different inteins is facilitated by a nomenclature based on amino acid position. The extein amino acid residue adjacent to the intein N-terminus is referred to as the −1 amino acid. The numbering then proceeds to the precursor protein N-terminus as described in Figure 1. The first C-extein residue is in the +1 position, the second is in the +2 position, and so on. The intein N-terminal residue is amino acid 1 with numbering proceeding in an increasing sequential manner to the end of the protein splicing element.

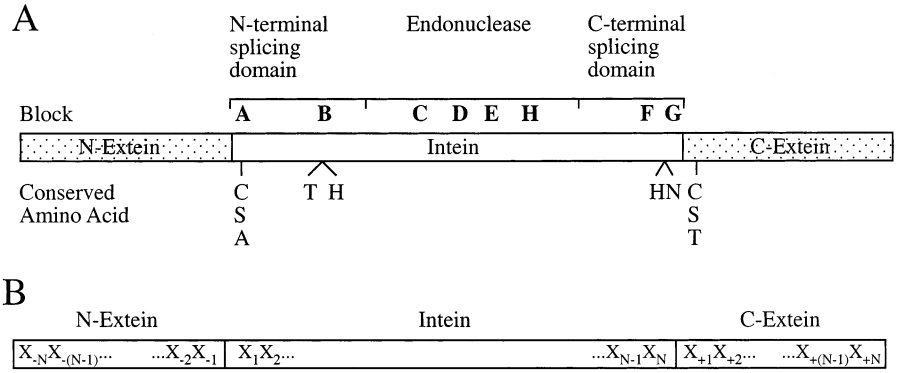


**Figure 1.** Intein motifs and numbering. (A) A protein splicing precursor contains the intein flanked by N- and C-exteins at its N- and C-termini, respectively. A number of primary sequence motifs have been discovered in protein splicing elements. Blocks A and B are present in the N-terminal splicing domain, while F and G are in the C-terminal splicing domain. Endonuclease-containing inteins also contain blocks C, D, E, and H. (B) An amino acid numbering system facilitates comparison between inteins. The first amino acid residue upstream of the intein is numbered −1 with numbering continuing upstream. The intein is numbered sequentially beginning with the N-terminal amino acid residue. The first C-extein residue is the +1 position with numbering proceeding toward the C-terminus.

Inteins can be divided into four basic classes: The maxi-intein, mini-intein, *trans*-splicing intein, and alanine intein. The maxi-inteins were the first to be discovered. They typically have an amino acid residue containing a nucleophilic side chain, such as a cysteine or serine, at their N-terminus and the conserved intein motifs blocks A and B (Figure 1). The N-terminal splicing domain is fused to a homing endonuclease domain. The homing endonuclease permits intein spread to inteinless genes via a homing mechanism.[19,20] After the endonuclease domain is the C-terminal splicing domain composed of intein motifs blocks F and G. Block G contains the conserved histidine and asparagine at the intein C-terminus and the downstream Cys, Ser, or Thr amino acid residue.

A mini-intein has the typical N- and C-terminal splicing domains; however, the endonuclease domain is not present. The first mini-inteins were created by removing the endonuclease domain at the gene level from the *Sce* VMA or *Mtu* RecA inteins.[21−23] Splicing in the absence of the endonuclease region demonstrated that protein splicing domains alone were sufficient for activity. Shortly thereafter, a naturally occurring mini-intein was reported to be present in the *Mycobacterium xenopi gyrA* gene.[24] Not surprisingly, the smallest known protein splicing element is a mini-intein, the *Mth* RIR1 intein,[25] and is only 134 amino acid residues in length. Despite its small size, this intein contains all the information necessary to splice in a foreign protein context.[26]

*Trans*-splicing inteins lack a covalent linkage between their N- and C-terminal splicing domains. Following the trend of the mini-inteins, *trans*-splicing was demonstrated using artificially split inteins before a naturally occurring form was discovered.[27−29] Interestingly, many artificially split inteins require a denaturant for optimal activity in vitro, while the naturally occurring *trans*-splicing intein, the *Ssp* DnaE intein, does not.[30] Furthermore, a recently described split *Sce* VMA1 intein also does not require a denaturant for activity.[31] The role of the denaturant is not fully understood; the two intein splicing domains may need to fold in the presence of one another, or the folded exteins may interfere with some aspect of the splicing reaction. The latter speculation is based on the observation that splicing will occur in the presence of a denaturant.[28] Whatever the reason, the *Ssp* DnaE intein splices in the absence of a chaotropic agent and allowed the first in-depth kinetic characterization of the protein splicing reaction (described below).

The *trans*-splicing *Ssp* DnaE intein may be the first reported instance in which the intein is vital to the host organism's survival.[4] It is found in the blue-green algae *Synechocystis* sp. PCC6803. The genes encoding the N- and C-terminal protein splicing domains fused to the appropriate gene fragments encoding the DnaE polymerase are located ca. 745 kilobases apart on the genome. Removal of the *trans*-splicing intein would presumably prevent the generation of full-length, active DnaE polymerase. This would almost certainly be detrimental because the DnaE polymerase is required for cell viability in other organisms. Liu and co-workers even speculated that *trans*-splicing is used as a regulatory element in *Synechocystis*,[4] although this has yet to be demonstrated. In contrast, other inteins can be removed from the host protein with no detectable change in activity. By extension, it is unlikely that removal of these inteins from an organism's genome is harmful.

The alanine inteins contain a naturally occurring N-terminal alanine residue.[18] Prior to the discovery of the alanine inteins it was thought that there was only one protein splicing mechanism, outlined in 1996.[32,33] This splicing process, which will be discussed in more detail in a later section, begins by converting the peptide bond upstream of the intein into a more reactive (thio)ester linkage. The conserved N-terminal amino acid residue of the first reported inteins was a cysteine or serine. The proposed protein splicing mechanism predicted that the side chain of this residue was involved in forming a (thio)ester bond. Demonstration that protein splicing and peptide bond cleavage at the N-terminal splice junction were entirely blocked by changing the conserved cysteine or serine to alanine supported this supposition. On the basis of these observations, when

the *Methanoccocus jannaschii* KlbA intein was found to contain a naturally occurring N-terminal alanine residue, it was assumed to be splicing incompetent.[18] Experimental evidence demonstrated that not only was the *Mja* KlbA intein active, but it spliced very efficiently in the test system utilized.[34,35] Furthermore, researchers were able to conclusively show that this intein bypassed the first step in the originally proposed splicing mechanism. Finally, these four categories are not necessarily exclusive. For example, it is possible for a mini-intein to also be an alanine intein or a *trans*-splicing intein.

## *III. Protein Splicing*

### A. Four-Step Mechanism

Intein splicing activity is vital for the proper functioning of many essential host proteins and, therefore, is often thought necessary for host organism viability. Despite the obvious importance of protein splicing, the mechanism was not conclusively elucidated until 1996. The final mechanism was the distillation of a wealth of previously accumulated information. Shortly after the discovery of inteins in *S. cerevisiae*[10,11] new protein splicing elements were reported in *Mycobacterium tuberculosis*[13] and *Thermococcus litoralis*.[36] Examination of the aligned sequences indicated the importance of the N- and C-terminal intein residues and the first downstream amino acid.[14] Mutagenesis studies further confirmed the pivotal role these residues played.[37]

The critical residues were a cysteine or serine at the intein N-terminus and a histidine−asparagine sequence at the C-terminus followed by a cysteine, serine, or threonine. The conservation of a histidine and either a cysteine or serine hinted at a mechanism similar to that of serine or cysteine proteases. Initially, three different protein splicing mechanisms were proposed, each explaining the importance of the conserved amino acid residues.[38−41] The impasse was broken with the characterization of a thermophilic intein from *Pyrococcus* sp GB-D.[38,39] It spliced in a foreign protein context, but the splicing could be slowed or stopped by lowering the temperature. This allowed researchers to demonstrate splicing of a purified precursor and the existence of a branched intermediate. It was determined that the first step in protein splicing was linear ester formation at the intein N-terminus.[32,33] Furthermore, C-terminal peptide bond cleavage was via succinimide formation of the final asparagine residue.[38,42]

These data fit one splicing model extremely well (Figure 2A). Protein splicing begins with an N−S or N−O acyl rearrangement to generate a (thio)ester bond between the intein N-terminal residue and the upstream extein residue. A trans(thio)esterification reaction occurs from attack of the hydroxyl or sulfhydryl group of the first residue downstream of the intein on this electrophilic reactive group. The resulting protein has two N-termini and is referred to as the branched intermediate. Succinimide formation by the intein C-terminal asparagine releases it from the two extein sequences, which are now linked by a (thio)ester bond. Splicing concludes with native peptide bond formation between the exteins by a spontaneous O−N or S−N acyl rearrangement.

### B. Three-Step Mechanism

The four-step splicing mechanism explained the importance of the conserved amino acid residues at the intein N- and C-termini. Other conserved amino acids, such as the block B or penultimate histidine (Figure 1), could be rationalized as assisting these junction residues. The identification by homology searches of a much larger number of inteins led to the finding of variants. Some of the new inteins had irregularities, such as ending in a glutamine instead of the conserved asparagine.[17] Even so, these differences could be reconciled with the accepted splicing mechanism. For example, both the conserved asparagine and the glutamine side chains can undergo a cyclization reaction coupled to peptide bond cleavage.

A dilemma arose with the discovery of inteins with an alanine in place of the conserved N-terminal serine or cysteine.[18] Either these inteins are inactive or their splicing proceeds by a different mechanism. It was suggested that an inactive intein would be retained if the host gene became dependent on it.[18] Although intein integration into host protein activity may occur, it was not necessary to explain the presence of the *Mja* KlbA intein. The *Mja* KlbA intein, which possesses a naturally occurring N-terminal alanine, was found to splice effectively in a foreign protein context.[34]

This brought attention to one of the originally proposed splicing mechanisms.[38] In this mechanism the N-terminal scissile peptide bond does not rearrange to a more reactive (thio)ester and is instead directly attacked by the downstream nucleophile (Figure 2B). Otherwise the mechanism was as previously reported. Interestingly, the *Mja* KlbA intein was able to catalyze the N−S acyl shift to a linear thioester intermediate when the N-terminal alanine was changed to a cysteine.[34] The researchers speculated that the ancestral intein used an N−S acyl shift. After a mutation changed the N-terminal cysteine to an alanine, the intein was still able to splice by direct downstream nucleophilic attack on the upstream peptide bond. Perhaps splicing efficiency was poor at first but would have increased with time in a selective environment. The demonstration of an alternative splicing mechanism offers the possibility that other splicing pathways may exist.

### C. Role of the Penultimate Histidine

The splicing mechanism explained the importance of the amino acids directly flanking the N- and C-terminal splice junctions. However, neither mechanism accounted for other conserved residues such as the block B threonine and histidine and the penultimate histidine. The penultimate histidine appears to assist C-terminal asparagine cyclization, leading to intein release from the C-extein. Mutating the penultimate histidine in the *Psp* polI intein abolished splicing.[32] Furthermore, the first two steps in protein splicing, the initial N−O acyl rearrangement and subsequent branch formation, still oc-

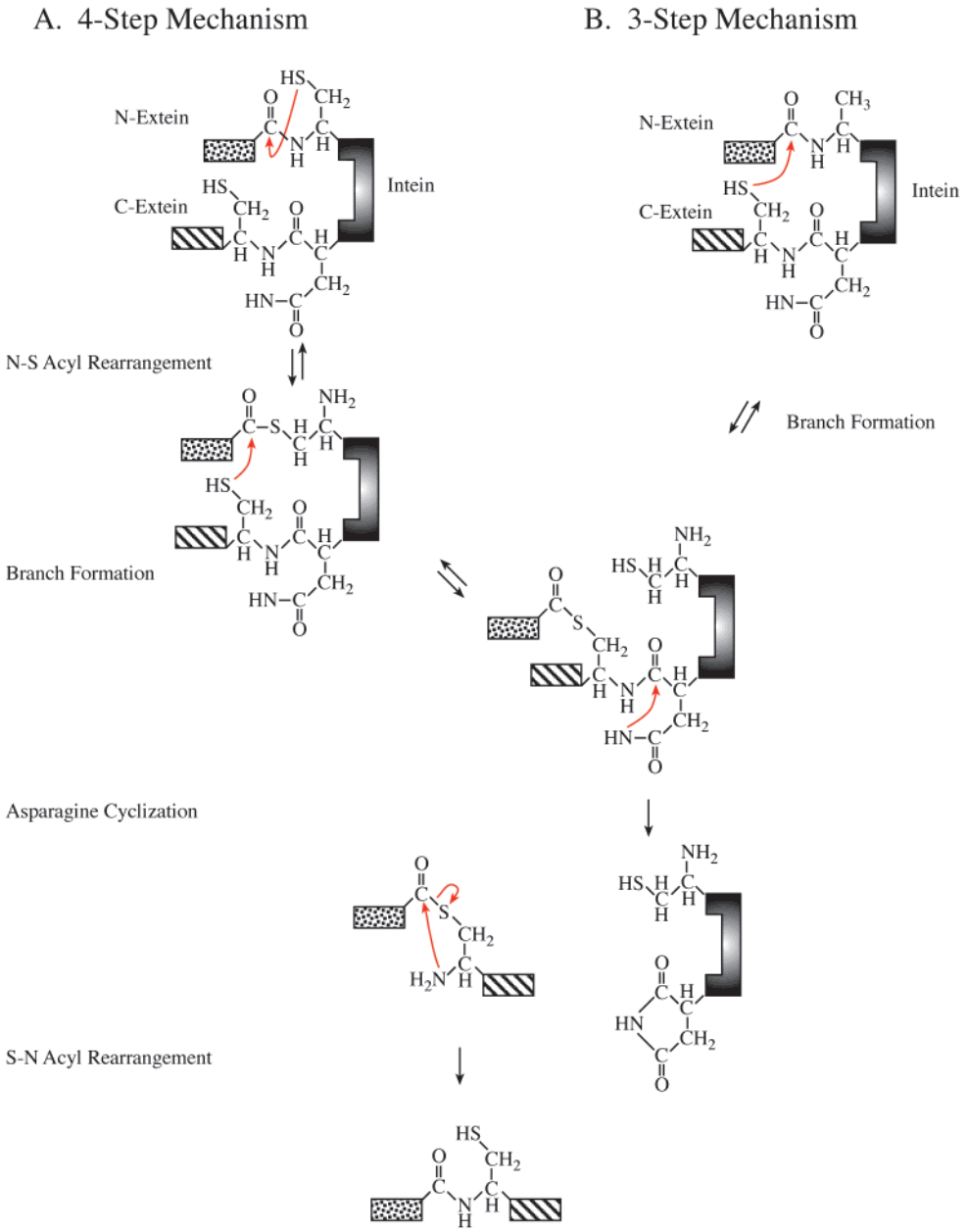


**Figure 2.** Protein splicing mechanism. (A) Four-step mechanism. The side chain sulfhydryl or hydroxyl of the intein N-terminal residue initiates an N−S or N−O acyl rearrangement. This generates a (thio)ester linkage between the N-extein and the intein. The downstream nucleophile, the side chain of Cys, Ser, or Thr, attacks the N-terminal (thio)ester in a trans(thio)esterification reaction to form the branched intermediate. Intein release from the branch is accomplished by cyclization of the intein C-terminal Asn coupled to peptide bond breakage. The (thio)ester bond between the N- and C-exteins is converted to a native peptide bond by a spontaneous S−N or O−N acyl rearrangement. (B) The three-step mechanism forgoes the initial N−S or N−O acyl shift. Instead the downstream nucleophile directly attacks the activated upstream peptide bond to form the branched intermediate. At this point the two splicing mechanisms follow the same path.

curred, but no C-terminal cleavage was detected. Structural data indicate the penultimate histidine protonates the amide nitrogen of the scissile peptide bond. The *Mxe* GyrA intein X-ray crystal structure has the penultimate residue, His197, hydrogen bonded to Asn198.[43] This would properly orient His197 to donate a proton from the N$\delta$ position to the amide nitrogen of the first C-extein residue. The three *Sce* VMA1 intein structures[44−46] report the close proximity of His453, the penultimate histidine, to the intein C-terminal Asn.

In comparison to the histidine found in the charge relay system of proteinases, the importance of the intein penultimate histidine is less obvious. This residue was changed in the *Sce* VMA1 intein to a number of other amino acids, and its effect on splicing was varied.[40] Changing the penultimate histidine to tryptophan or glycine inhibited splicing. Alternatively, leucine or valine did not abolish splicing even though they could not obviously participate as a proton donor/acceptor. Furthermore, changing the penultimate histidine to glutamine in the *Sce* VMA1 intein coupled N- and C-terminal cleavage.[47] Examining the nature of this effect may give some insight into how inteins regulate the steps in the splicing pathway.

A number of putative inteins naturally lack the penultimate histidine residue. If this residue is

needed for splicing, then these inteins are either inactive or have substituted other amino acid residues for the histidine activity. Five of these inteins were examined, and four were found to splice: the *Ssp* DnaE, *Mja* PEP, *Mja* Rpol A′, and *Mja* KlbA inteins.[4,34,48,49] Changing the penultimate residue to a histidine improved splicing extent in the *Mja* PEP intein, while the *Mja* Rpol A′ and *Mja* KlbA inteins had decreased splicing activity.[34,48] There are conflicting reports on the effect of changing the *Ssp* DnaE intein penultimate residue to a histidine.[50,51] Data interpretation is complicated by the unknown effect that non-native amino acids have on the geometry of the tightly packed splice junction residues. For example, when the *Mja* KlbA intein penultimate residue, Ser167, was changed to histidine, there was less spliced product but more C-terminal cleavage.[34] This could be due to an increased asparagine cyclization efficiency, leading to C-terminal cleavage competing with splicing, or the histidine may have distorted the splice site conformation and interrupted the highly coordinated protein splicing steps.

Despite reports that the penultimate histidine can be changed without preventing protein splicing, the cumulative data indicate that this residue is assisting in asparagine cyclization. Studies performed to date have often looked for gross changes, for example, a detectable change in splicing or cleavage activity following an overnight incubation. The histidine may instead be increasing the splicing rate, but may not be absolutely required for activity. Kinetic characterization of histidine mutants may provide a clearer picture by detecting significant, but less drastic, changes in splicing activity.

## D. Conserved Block B Threonine and Histidine

The block B motif of protein splicing elements contains a conserved Thr and His typically separated by two amino acid residues. The block B histidine shows less variability than the penultimate histidine and, perhaps surprisingly, than the catalytically defined N-terminal cysteine. Only the Asn and Cys, Ser, or Thr residues at the C-terminal splice junction are more conserved. Despite this fact, far fewer studies have been published on the role of the block B histidine in protein splicing.

Changing the block B histidine to leucine or asparagine in the *Sce* VMA1 intein abolished[52] or decreased,[46] respectively, splicing activity in *E. coli*. Western blot analysis showed the presence of a low molecular weight band at the position expected for the product of C-terminal peptide bond cleavage. Furthermore, the lack of any observable N-terminal cleavage product implicated the block B histidine in the first step of protein splicing, the N−S acyl rearrangement. A block B mutant of the *Sce* VMA1 intein was also tested for activity in an in vivo complementation system.[52] In this experiment the yeast required an active vacuolar $H^+$-ATPase for growth under selective conditions. Very weak growth indicated a low level of activity that arose from splicing of the mutant intein.[52] Therefore, protein splicing can occur in the absence of the block B histidine, although at a significantly impaired efficiency. *Trans*-splicing activity is also abolished in the *Ssp* DnaE intein when the block B His is changed to alanine,[53] indicating that it is of general importance.

The block B histidine, like the penultimate histidine, may be a component of a charge relay system. The high conservation of the block B histidine and the dramatic decrease in intein splicing activity when this residue is mutated mirrored the behavior of subtilisin, a protease utilizing a catalytic triad.[54] The first intein crystal structure reported the close proximity of the block B histidine and the N-terminal cysteine,[44] supporting its role in proton abstraction from the cysteine sulfhydryl group. This would activate the sulfhydryl group and facilitate the N−S acyl shift. However, interpretation of events at the intein termini is complicated by the lack of extein amino acid residues in this structure. Later X-ray crystal structures of the *Mxe* GyrA and *Sce* VMA1 inteins with extein residues predicted an alternate interaction.[43,45,46] The imidazole ring of the block B histidine was in position to protonate the amide nitrogen of the N-terminal scissile peptide bond. This promotes breakdown of the tetrahedral intermediate formed during downstream nucleophilic attack on the upstream (thio)ester bond (Figure 2, branch formation).

The alanine inteins provide further insight into the role of the block B histidine. This conserved residue is present in the nine currently reported alanine inteins.[3] Presumably, these inteins do not use the N−S acyl shift as part of their splicing mechanism. If the block B histidine assists in the N−S acyl shift, then in the alanine inteins it is a residual amino acid not yet substituted by random mutation. Perhaps, in both the alanine and common inteins this histidine promotes breakdown of the tetrahedral intermediate.[43,45,46]

Also present in block B is the highly conserved Thr. Mutating this residue alone or in conjunction with the block B histidine abolished any detectable protein splicing activity of the *Ssp* DnaE[53] or *Sce* VMA1[52] inteins. The block B histidine in the *Mxe* GyrA intein crystal structure was in contact with the amide nitrogen of the scissile peptide bond.[43] This interaction was postulated to stabilize the unusual *cis* conformation of the scissile peptide bond. The energetically unfavorable *cis* bond lowers the activation energy for nucleophilic attack on the carbonyl carbon. Furthermore, the block B threonine hydroxyl and the N$\delta$ of Asn74 form an oxyanion hole that could stabilize the resulting tetrahedral intermediate.

The *Sce* VMA1 intein crystal structures containing extein residues did not have a *cis* scissile peptide bond. Instead, this bond was either in a strained[45] or a normal *trans* conformation.[46] The strained *trans* peptide bond had a $\tau$ angle of 100°, 10° from ideal. This energetically unfavorable conformation was stabilized by hydrogen bonding between the activated N-terminal and Ile434 peptide bonds. Furthermore, this intein contained a $Zn^{2+}$ bound in the active site. The $Zn^{2+}$ is known to inactivate the intein, although the mechanism is unknown. Therefore, the distorted

*trans* peptide bond is either an artifact or a splicing conformation frozen by $Zn^{2+}$ binding. This would not be the first report of a distorted *trans* peptide bond facilitating bond cleavage. It is reminiscent of the glycosylasparaginase self-cleavage mechanism.[55] In this mechanism, the strained *trans* conformation of the scissile peptide bond drives ester formation by an N−O acyl shift. Subsequent ester hydrolysis releases an N-terminal threonine required for glycosylasparaginase activity. Similarly, the intein strained *trans* peptide bond may assist in the initial N−S or N−O acyl rearrangement. Alternatively, the N−S or N−O acyl shift of a typical *trans* peptide bond may be facilitated by deprotonation of the intein N-terminal nucleophile. Crystal structure data indicated that the *Sce* VMA1 N-terminal nucleophile may be activated by the upstream Gly carbonyl oxygen.[46] The Cys1 thiolate anion then attacks the −1 residue carbonyl carbon to generate a tetrahedral intermediate which has a thiazolidine structure. However, this intein was mutated to slow splicing. Therefore, it is difficult to determine whether these structures are part of the actual splicing mechanism or non-native activities. Further experimental analyses will need to be performed to answer these questions and determine the correct mechanism.

## E. Divalent Transition-Metal Cation Influence on Protein Splicing

Surprisingly, a recent *Sce* VMA1 intein crystal structure reported $Zn^{2+}$ bound in the active site. Zinc ion coordination was through Glu80 and His453 of the intein, the first C-extein residue, and a water molecule.[45] Atomic absorption spectroscopy revealed a 1:1 stoichiometry of $Zn^{2+}$ to intein. The $Zn^{2+}$ binding affinity must be high because it copurified with the intein even though there was no detectable $Zn^{2+}$ in the crystallization or storage buffers. The purpose, if any, of the $Zn^{2+}$ was not clear from the crystal structure. Metalloproteases utilize divalent metal cations to assist in peptide bond cleavage. However, the bond angles of the intein-bound $Zn^{2+}$ were more indicative of a structural role. If it served a structural role, then there may be conformational differences between inteins with and without $Zn^{2+}$. The $Zn^{2+}$-containing *Sce* VMA1 intein and the *Mxe* GyrA intein lacking $Zn^{2+}$ did not vary significantly in the backbone structure. The largest change[43,45] was that the $\chi_2$ of the penultimate histidine imidazole ring differed by 136°.

The $Zn^{2+}$ was not involved in large structural perturbations, but it was in contact with two key catalytic amino acid residues, the penultimate histidine and the +1 Cys. In fact, it was in position to block protein splicing by tying up the +1 Cys side chain and preventing its involvement in transthioesterification. Two groups subsequently demonstrated that $Zn^{2+}$ inhibits protein splicing of two different inteins, the *Ssp* DnaE and the *Mtu* RecA inteins.[51,53] As described previously, the *Ssp* DnaE intein is a naturally occurring split intein. The N- and C-terminal splicing domains can be expressed and purified separately and splicing activity reconstituted by mixing the two domains in vitro.[30] Splicing was inhibited when $Zn^{2+}$ and the two intein splicing domains were mixed simultaneously. Total inhibition was observed at millimolar $Zn^{2+}$ concentrations, while splicing decreased ca. 85% at 10 $\mu$M. Even 1 $\mu$M $Zn^{2+}$ had a small but significant effect on splicing. Interestingly, the *Ssp* DnaE intein has an alanine at the penultimate position and lacks an equivalent Glu80. If $Zn^{2+}$ is binding to this intein, then other residues must be involved in chelation. Amino acid residue Glu85 or Glu86 may supply the ligand from the *Ssp* DnaE intein that Glu80 supplies from the *Sce* VMA1 intein. Other divalent metal cations were evaluated, but only $Cd^{2+}$ approached $Zn^{2+}$ in inhibiting protein splicing activity.[51,53] Other divalent metal cations, $Co^{2+}$, $Ni^{2+}$, $Mg^{2+}$, and $Ca^{2+}$, had little or no detectable effect on splicing.

The *Mtu* RecA intein provided a unique opportunity to evaluate the $Zn^{2+}$ effect on both *trans*- and *cis*-splicing.[51] The *Mtu* RecA intein occurs naturally as a maxi-intein. However, the $Zn^{2+}$ studies utilized a previously generated *Mtu* RecA mini-intein. The new mini-intein was further divided into its N- and C-terminal splicing domains to create a *trans*-splicing *Mtu* RecA intein. Unlike the naturally split *Ssp* DnaE intein, the artificially split *Mtu* RecA intein requires a denaturation/renaturation step to *trans*-splice.[27] Despite the differences between these two inteins, 2 mM $ZnCl_2$ decreased the *Mtu* RecA intein *trans*-splicing activity by >90%.

Experiments with the *Mtu* RecA mini-intein demonstrated that $Zn^{2+}$ could inhibit a *cis*-splicing intein. Typically, the *cis*-splicing reaction is rapid and uncontrollable, making isolation of the full-length precursor difficult. This was circumvented by fusing the *Mtu* RecA intein between two hexahistidine tags and purifying the resulting precursor protein from inclusion bodies.[51] Protein splicing activity could be reconstituted in vitro by refolding the aggregated inteins in serial dilutions of urea. Intein refolding in the presence of $Zn^{2+}$ inhibited splicing and N-terminal cleavage activity. This finding still leaves unanswered whether $Zn^{2+}$ can bind and interfere with a folded intein or whether divalent metal cation binding must occur concomitantly with protein splicing domain association. It also remains to be determined whether $Zn^{2+}$ cations interfere or modulate the activity of inteins in vivo. This may be unlikely because the free $Zn^{2+}$ concentration in *E. coli* cells was estimated to be in the femtomolar range.[56] Incredibly, this is less than one free $Zn^{2+}$ ion per cell. Therefore, intracellular $Zn^{2+}$ would not affect protein splicing unless the intein effectively competed for bound $Zn^{2+}$. It is likely that the effect of $Zn^{2+}$ and $Cd^{2+}$ on protein splicing is an interesting but nonphysiological phenomenon.

## F. Kinetic Characterization

The kinetic analysis of enzyme reactions is a powerful tool for elucidating catalytic mechanism and catalytically relevant amino acid residues. Many proteins that facilitate chemical reactions behave as catalysts and interact with their substrates in a conceptually similar manner. This fact permitted simplifying assumptions to be made in the derivation

of expressions such as the Henri−Michaelis−Menten equation.[57] Analyses using these expressions have shaped our understanding of enzymes, but it is important to remember the underlying criteria on which they are based. One criterion is that an enzyme undergo turnover. Importantly, there are a growing number of proteins that facilitate a chemical process but do not display turnover. These include the recently described hedgehog proteins, glycosylasparaginases, and inteins that undergo autoproteolysis or splicing. As discussed previously,[58] an intein excises itself from a protein precursor only once. The issue may seem minor; after all the intein is spatially arranging reactive groups to lower the activation energy of protein splicing and therefore assist the reaction in an enzyme-like fashion. It is, in actuality, an important consideration when kinetic rate constants are determined.

An in-depth kinetic analysis would complement and expand upon the X-ray crystallographic and mutagenic intein studies. X-ray crystallography gives a snapshot of the protein. Ideally, the structure of protein splicing intermediates and their rate of interconversion would be determined. For example, one of the *Sce* VMA1 intein X-ray crystal structures, representing a $Zn^{2+}$-inactivated complex, found the +1 residue of the C-terminal splice junction positioned 9 Å from the N-terminal scissile peptide bond.[45] If this represents a true splicing conformation, then understanding events that bridge this 9 Å gap, resulting in transthioesterification, may allow the rational control of protein splicing at this level.

Inteins have been extensively characterized by mutagenesis, leading to the determination of catalytically important amino acid residues. These studies usually involve assaying for drastic changes in splicing or cleavage activity. The coupling of kinetic and mutagenic analyses may show the relevance of amino acid residues with subtle effects on protein splicing. A serious limitation of intein kinetic studies is the need to controllably initiate the splicing reaction at time zero. This is no trivial matter, and the final elucidation of the protein splicing mechanism itself awaited the characterization of a thermophilic intein. Researchers could arrest this intein at different splicing steps by changing the reaction temperature.

In fact, the *Psp* GB-D pol1 intein from a thermophilic archaeon was one of the first protein splicing elements for which the splicing and N-terminal peptide bond cleavage half-lives were determined. The intein was sandwiched between maltose binding protein (MBP) and a fragment of *Dirofilaria immitis* paramyosin.[39] In this foreign protein context and at 37 °C the *Psp* GB-D pol1 intein spliced with a half-life ($t_{1/2}$) of 30 min. Assuming a first-order reaction, this would correspond to a rate constant for splicing of $3.8 \times 10^{-4}$ $s^{-1}$. The N-terminal cleavage $t_{1/2}$ of the wild-type intein was not determined, but a mutant form (Ser1Cys) had a $t_{1/2}$ of 20 min. This corresponds to a $k_{obs}$ of $5.8 \times 10^{-4}$ $s^{-1}$. The time-dependent cleavage reactions of two other mutant inteins were also reported. These inteins were modified to abolish splicing but permit N-terminal cleavage. Converting the reported N-terminal cleavage half-lives for the mutant *Sce* VMA1 (at pH 8)[47] and artificial *Mtu* RecA mini-inteins[59] results in first-order rate constants of $1.9 \times 10^{-3}$ and $3.5 \times 10^{-5}$ $s^{-1}$, respectively.

These important studies yielded information on pH and mutational effects. However, the intein progression along the splicing or cleavage pathway at time zero was not known. Even for the temperature-dependent *Psp* GB-D pol1 intein it was difficult to determine whether 100% of the intein was at step 1 in the splicing pathway when the temperature was increased at time zero. It is likely that the cold simply slowed the reaction, leaving a heterogeneous population. The same was true for intein mutants that no longer spliced but underwent peptide bond cleavage at the intein N-terminus upon thiol addition. At time zero, is the scissile bond an amide or has it started along the reaction pathway and already formed a thioester? This situation became tractable with the discovery of the *Ssp* DnaE intein, the naturally occurring *trans*-splicing intein.

The utility of *trans*-splicing is that the N- and C-terminal splicing domains can be mixed at a known time. Before mixing the intein fragments are not progressing along the splicing pathway. The evidence for this comes from the fact that separated N- and C-terminal domains of the *Ssp* DnaE intein do not undergo any detectable proteolysis or splicing in the presence or absence of a thiol reagent (Martin et al., New England Biolabs, Inc., unpublished results). Shortly after mixing, the intein fragments cleave or splice under the same conditions. Although *trans*-splicing with artificially split inteins was reported before the discovery of the *Ssp* DnaE intein, their requirement for treatment with a chaotropic salt complicated kinetic studies. The *Ssp* DnaE intein does not need this treatment. Despite the advantage of controllable initiation, *trans*-splicing adds a complicating step to the splicing pathway. Specifically, the two splicing domains must bind in the correct orientation to initiate the reaction. Indeed, the association must be considered when the $k_{obs}$ is determined for each step in the splicing pathway. It may be, however, that the requirement for association of the two splicing domains is not unique to the *trans*-splicing reaction. The two splicing domains of *cis*-splicing inteins may fold independently and then associate to initiate protein splicing. This may explain why artificially split inteins require a denaturant, but not necessarily its removal, for *trans*-splicing activity.[28] It may unfold the extein sequences and permit the folded intein fragments to come together. Of course, covalently attached splicing domains would facilitate association compared with the split inteins.

Specific *Ssp* DnaE intein splicing steps can be studied by varying the reaction conditions. The N−S acyl shift and C-terminal asparagine cyclization were followed by DTT-induced N- and C-terminal cleavage, respectively. Splicing rates were determined in the absence of a thiol reagent. A summary of the observed rate constants[60] is shown in Figure 3. The $k_{obs}$ for DTT-mediated peptide bond cleavage at the N-terminal splice junction was $1.0 \times 10^{-3}$ $s^{-1}$. The $k_{obs}$

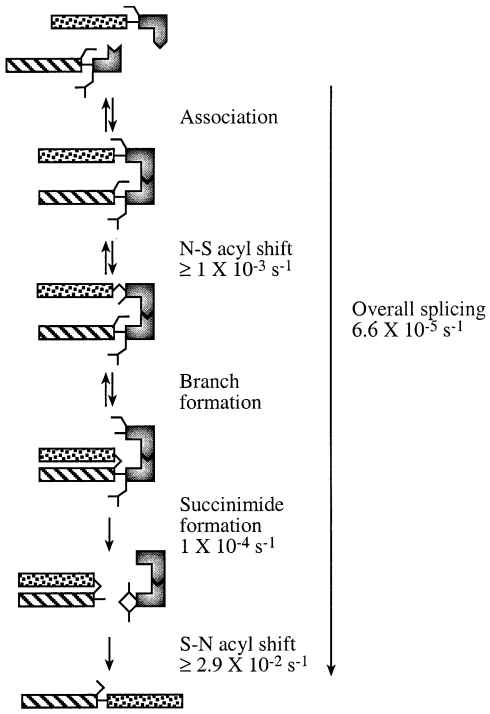


**Figure 3.** Kinetic analysis of splicing and splicing steps. The naturally occurring *trans*-splicing *Ssp* DnaE intein was used for the kinetic characterization of protein splicing. Experimental limitations precluded determining the rates of association and branch formation.

for N-terminal cleavage is a function of the N−S acyl shift rate and the rate of DTT-mediated nucleophilic attack on the newly formed thioester. Therefore, the N−S acyl shift is at least as fast as the rate of DTT-induced N-terminal cleavage. The $k_{obs}$ for DTT-induced N-terminal cleavage of the *Ssp* DnaE intein was comparable with the rates of the *Sce* VMA1[47] and mutant Ser1Cys *Psp* GB-D[39] pol1 inteins and ca. 30-fold faster than the *Mtu* RecA mini-intein[59] rate.

The asparagine cyclization rate, step 3 of the splicing mechanism, was determined by following the time-dependent formation of C-terminal cleavage product. Unexpectedly, scissile peptide bond cleavage at the *Ssp* DnaE intein C-terminal splice junction did not occur until cleavage at the N-terminal splice junction was initiated. Coupling these events may orchestrate the complex splicing reaction to minimize side reactions. Even though the *Ssp* DnaE intein couples N- and C-terminal peptide bond cleavage, this does not appear to be a universal intein property. In fact, most inteins tested undergo C-terminal cleavage in the absence of N-terminal cleavage.[24,26,47,61−63] Although most wild-type inteins do not absolutely couple N- and C-terminal cleavage, a His453Gln mutation generates this behavior in the *Sce* VMA1 intein.[47]

The $k_{obs}$ for the overall splicing reaction was determined to be $6.6 \times 10^{-5}$ $s^{-1}$. Therefore, none of the steps in the splicing reaction for which a $k_{obs}$ was reported are rate-determining steps. There were two steps for which rate constants were not determined, intein fragment association and transthioesterification. Association of the splicing domains appears to be very fast relative to the other steps in protein splicing and is likely not to be rate determining. In addition, once the N- and C-terminal splicing domains bind, the off rate appears to be extremely slow. This is based on the lack of turnover for the N-terminal *trans*-cleavage constructs. In contrast, the *trans*-cleaving $\delta$-ribozyme, the RNA equivalent of intein-mediated *trans*-cleavage, binds its cognate RNA, induces cleavage, and dissociates and then repeats the process relatively rapidly.[64]

Recently, an artificially split *Sce* VMA1 intein was reported to *trans*-splice without the need for a denaturant.[31] Comparable to that of the *Ssp* DnaE intein, the splicing reaction kinetics could be fit to a first-order equation and yielded a $k_{obs}$ of $1.9 \times 10^{-4}$ $s^{-1}$. The splicing reaction utilized dimerization domains fused to the intein fragments. This implies a lower affinity interaction between the *Sce* VMA1 intein N- and C-terminal splicing domains in this construct than previously observed with the *Ssp* DnaE intein.

Measuring the N- and C-terminal splicing domain affinity and the association and dissociation rates would facilitate the discovery of amino acid residues involved in splicing domain dimerization. Studies with the *Ssp* DnaE intein found that deleting up to 29 amino acid residues from the C-terminus of the N-terminal splicing domain still allowed association. Furthermore, three amino acid residues could be removed from the N-terminus of the C-terminal splicing domain without abolishing binding.[53] These limits may reflect the importance of subsequent amino acid residues in association or indicate a minimum length for proper intein folding. Investigation of the block F residues demonstrated that no single amino acid change abolished binding; however, splicing was no longer detected.[53] This was the first systematic study of block F and indicates that these conserved amino acid residues are involved in catalysis or folding and not in binding.

Interestingly, three of the four inteins examined have similar N-terminal cleavage rates. Furthermore, the *Ssp* DnaE and *Sce* VMA1 inteins have similar $k_{obs}$ values for *trans*-splicing. It is possible that most protein splicing rates are optimized to have no impact on the host organism. However, in certain instances the rate of protein splicing may be co-opted by the cell as a regulator of protein function. This point has been proposed as a role for the *Ssp* DnaE intein.[4] Selective pressure would likely work on the entire splicing pathway. Even so, in a limited sampling 75% of the inteins studied had N-terminal cleavage rates that varied by less than 2-fold. This is particularly striking because changing the extein sequences flanking the same intein can significantly alter its cleavage activity[47,61,65] and rate (Nichols et al., New England Biolabs, Inc., unpublished results).

The *Ssp* DnaE intein has permitted kinetic characterization of protein splicing steps, an important start, but the real potential is in future investigations. The published rates provide a marker to evaluate mutant forms of the *Ssp* DnaE intein. In addition to altering the intein itself, researchers can investigate in more detail the effect of changing the extein sequences. This will lead to an expanded understanding of the protein splicing mechanism and

intein evolution. Comprehensive kinetic characterization of other inteins, such as the *trans*-splicing *Sce* VMA1 intein, will shed light on how protein splicing elements interact with their host protein and organism. It is well documented that parasites can have a profound effect on a host organism's evolution. If inteins are truly a form of parasitic DNA, perhaps these intriguing molecules are actively involved in determining the evolutionary direction of its hosts.

## *IV. Applications*

### A. Protein Purification

Investigations into the splicing mechanism showed that mutating particular catalytic residues could disrupt the tightly coupled splicing process. In particular, certain amino acid changes resulted in peptide bond cleavage at the N- or C-terminal splice junction.[32,33,40,52] Therefore, one of the first intein-based applications was in protein purification as a self-cleaving affinity tag. Controllable N-terminal peptide bond cleavage was achieved by mutating the conserved intein C-terminal asparagine. This abolished protein splicing and C-terminal cleavage activity but still allowed peptide bond cleavage at the N-terminal splice junction.[66] Cleavage was induced by nucleophilic attack on the thioester formed in inteins containing an N-terminal cysteine residue. Fusion of the chitin binding domain (CBD) from *Bacillus circulans*[67] to the modified *Sce* VMA1 intein created an intein−tag that could be purified on chitin beads (Figure 4A). The gene encoding the target protein is cloned upstream of the intein−tag gene using a multiple cloning site.

Originally, protein expression was under the control of the tac promoter, but the current commercially available kits (the IMPACT-CN and IMPACT-TWIN kits) utilize the tightly regulated T7 promoter.[68] Following transcription and expression in *E. coli*, the target−intein−tag fusion protein is released from the bacterial cell by lysis. Cell debris is pelleted, and the supernatant is applied to a chitin column. Unbound proteins are washed from the column. The target protein is separated from the column-bound intein−tag by incubating the resin with hydroxylamine or, more commonly, DTT. The released target protein is then eluted from the column. The system requires no expensive and potentially problematic proteases, and intein-mediated cleavage can be designed to leave no extra amino acid residues on the target protein.

Certain proteins have higher expression levels if a highly expressed polypeptide is fused to its N-terminus. To create an N-terminal intein−tag, it was necessary to have controllable peptide bond cleavage at the intein C-terminal splice junction. A method for controllable cleavage was discovered during studies on the *Sce* VMA1 intein penultimate His. Changing the penultimate His to a Gln coupled C-terminal peptide bond cleavage to DTT-induced N-terminal peptide bond cleavage.[47] Placement of the chitin binding domain into the endonuclease region permitted affinity purification.[69] The intein−tag is fused to the N-terminus of the target protein (Figure 4B). After IPTG-induced expression, the target−intein−tag fusion protein is immobilized on a chitin resin. Addition of DTT induces release of a short peptide from the intein N-terminus. This stimulates peptide bond cleavage between the intein−tag and target protein. Eluted fractions contain both the target protein and the small peptide, which can be removed by dialysis. Recently, a method was developed in which GFP was incorporated into the intein−tag as a rapid marker of protein expression and solubility.[70] This will allow high-throughput screening for optimally expressed proteins from a library of genes.

Inteins can also be modified to undergo C-terminal splice junction cleavage in a pH- and temperature-dependent fashion (Figure 4C). This has allowed the purification of proteins possessing an N-terminal residue other than Met, for example, with an N-terminal cysteine for use in intein-mediated protein ligation (discussed below). Three protein splicing elements were modified for this purpose, the *Mth* RIR1,[26] *Ssp* DnaB,[62] and *Mxe* GyrA inteins.[61] A forth intein, the *Mtu* RecA mini-intein, was selected for C-terminal cleavage activity using a potent thymidylate synthase reporter system.[63] This intein in conjunction with maltose binding protein as an affinity tag was later used to purify human acidic fibroblast growth factor.[71]

### B. Protein Ligation

An interesting protein purification intermediate using a thiol-inducible intein−tag is the generation of a transient C-terminal thioester on the target protein. Radiolabeled cysteine was ligated to the thioester-tagged protein through a native peptide bond[66] using chemistry described previously.[72,73] However, the extent of ligation was low. Ligation extents were increased to >90% by changing the thiol reagent used to induce intein cleavage from DTT to 2-mercaptoethanesulfonic acid (MESNA) or thiophenol.[74−76] MESNA is often preferred over thiophenol because of its solubility in water and innocuous odor. Furthermore, a synthetic peptide with an N-terminal cysteine was ligated to the thioester-tagged protein in place of the radiolabeled cysteine. This technique was termed intein-mediated protein ligation (IPL) or expressed protein ligation (see Figure 5). Initially, protein ligation was used to purify cytotoxic proteins,[74] to introduce fluorescent probes into a protein sequence,[75] or to determine a region of protein−protein interaction.[76] A variation on the ligation technology is to use ammonium bicarbonate in place of a cysteine or peptide to generate amidated peptides.[77] A number of pharmaceutically relevant peptides require carboxy amidation for biological activity, making this technique particularly useful for studies of potential therapeutic agents.

Intein-mediated protein ligation has also permitted protein immobilization and antibody labeling. Immobilization of proteins in microarrays offers the potential of elucidating protein function or interactions using high-throughput screening assays. Protein immobilization through amino acid side chains

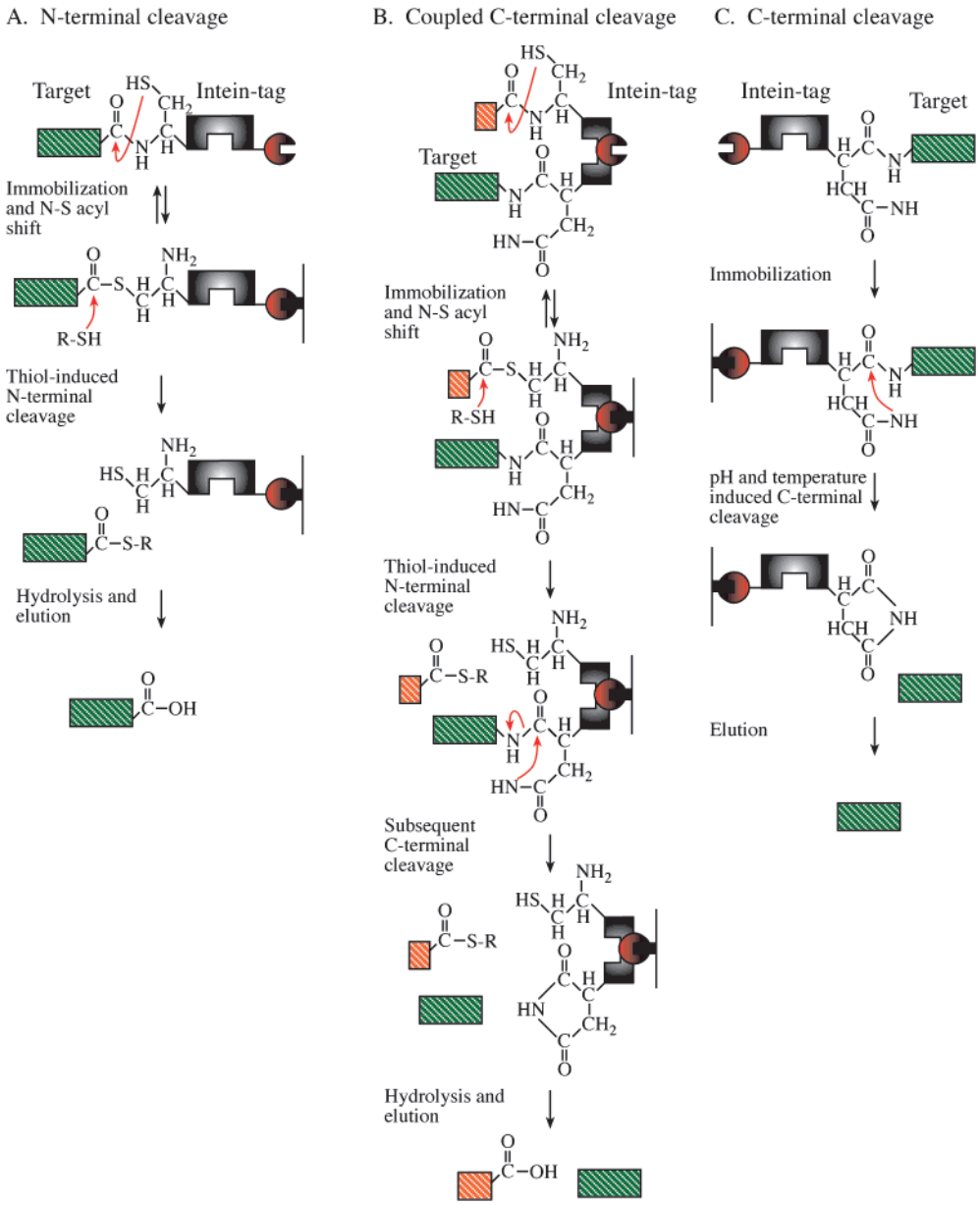


**Figure 4.** Intein-based protein purification. (A) Thiol-induced N-terminal cleavage. The intein−tag is composed of an intein modified to undergo N-terminal cleavage and an affinity tag, typically the CBD. The intein−tag is fused to the C-terminus of the target protein at the gene level. Following expression and cell lysis, the fusion protein is immobilized on an affinity column. The target is released by thiol-induced cleavage of the bond between the target protein and the intein. The pure target is subsequently eluted from the column. (B) Coupled C-terminal cleavage. Changing His 453 of the *Sce* VMA1 intein to a Gln results in an intein that cleaves at its C-terminus only after thiol-induced N-terminal cleavage. The intein−tag has the affinity tag inserted into the protein splicing element. This type of intein−tag allows purification of target proteins with N-termini composed of non-Met amino acid residues, with the exception of Cys or Ser. (C) C-terminal cleavage. A number of inteins have been altered to undergo pH- and temperature-dependent cleavage at their C-terminus. The intein−tag typically contains a chitin binding domain or maltose binding protein as the affinity group. These constructs permit the isolation of proteins with a non-Met N-terminus including an N-terminal Cys or Ser.

results in the protein being linked to a solid support in random orientations. Many of these orientations are nonproductive. Attaching a binding or reactive group to the target protein site specifically allows immobilization in a single orientation.[78] IPL was applied to site specifically attach a biotin moiety to three proteins that were subsequently immobilized on avidin-functionalized glass slides.[79] The utility of IPL is that the biotin moiety can be replaced by a variety of functionalities, depending on the experimental requirements. Furthermore, IPL was used to label a single-chain antibody for biochip screening.[80] Antibody expression in *E. coli* often results in the formation of inclusion bodies. Performing the ligation reaction in 4 M urea permitted the site-specific labeling of a solubilized antibody. The labeled antibody displayed the expected binding preference for neutravidin compared to streptavidin and avidin in a biochip experiment.

Expansion of ligation technology to the fusion of two bacterially expressed proteins required that one of the reactants contain an N-terminal cysteine. The requisite N-terminal cysteine can be released during purification with an intein modified to undergo

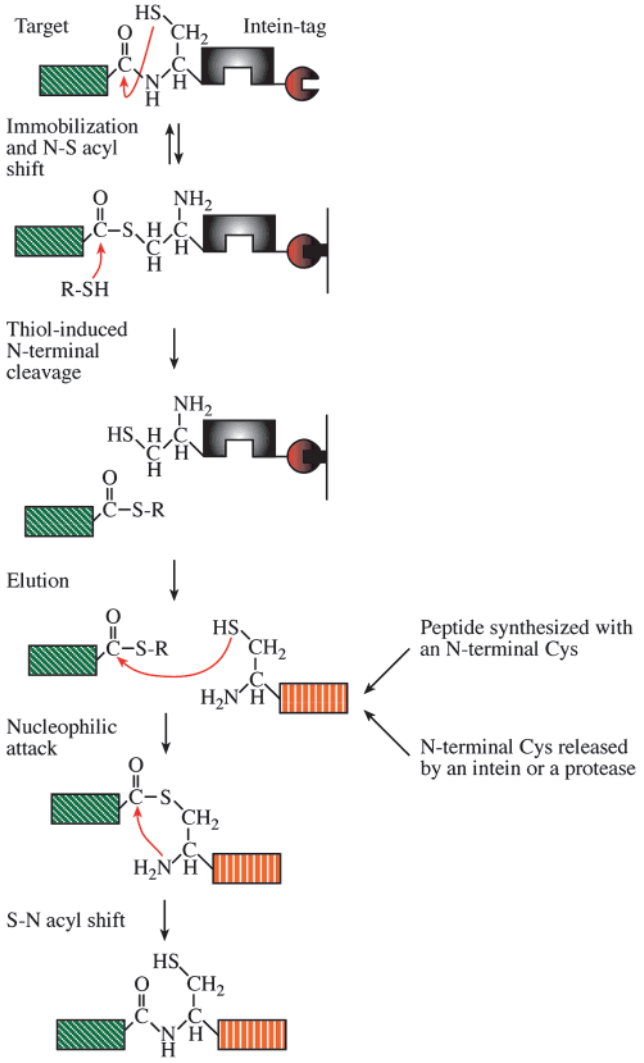


**Figure 5.** Intein-mediated protein ligation. IPL takes advantage of the C-terminal thioester generated on a target protein following thiol-induced N-terminal cleavage. This thioester-tagged protein reacts with a protein or peptide with an N-terminal cysteine to initially attach the proteins through a thioester bond. As in protein splicing, this thioester spontaneously rearranges to generate a native peptide bond between the reacting species.

cleavage at the C-terminal splice junction[26,65] or by protease treatment.[81,82] The ligation of two polypeptides permits the labeling of specific protein regions for NMR analysis.[83] Furthermore, coupling these technologies resulted in proteins with both an N-terminal cysteine and a C-terminal thioester. This species undergoes either an inter- or intramolecular reaction, leading to multimeric or cyclic protein species, respectively.[65,84,85] Polymerization of a repeating unit would facilitate the study and development of designer biopolymers with novel properties. Merging the principles of solid-phase synthesis and protein ligation permitted the controlled assembly of a protein species.[86] This technique, termed solid-phase expressed protein ligation (SPPL), was used to generate a dual-labeled biosensor for the study of protein−protein interactions in the nonreceptor protein tyrosine kinase, c-Abl. Importantly, SPPL allows the site-specific incorporation of probes into almost any protein sequence.

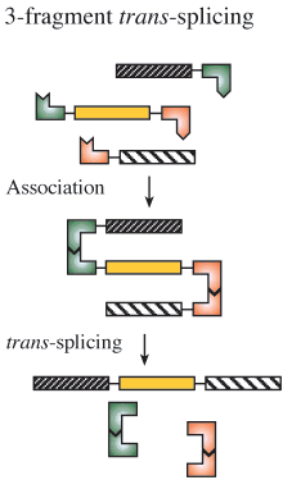


**Figure 6.** Three-fragment *trans*-splicing. The genes for two *trans*-splicing inteins are fused to the appropriate three gene fragments. The three fusion proteins are purified following expression. The intein splicing domain selectivity results in the correct full-length product.

## C. *Trans*-Splicing

The ligation of protein segments can also be accomplished by *trans*-splicing.[27−31,87] Typically, the *trans*-splicing reaction proceeds more efficiently at lower reactant concentrations than IPL, making it a compelling alternative. The intein fragments associate and bring the reactants together, whereas IPL reactants lack this favorable binding. Segmental isotopic labeling using *trans*-splicing has been demonstrated for an RNA polymerase α subunit and MBP.[88,89] Also, the central portion of MBP was labeled by simultaneous three-fragment *trans*-splicing[90] (Figure 6).

The three-fragment *trans*-splicing experiment illuminates certain intein properties. The respective N- and C-terminal splicing domains of the two inteins *Pfu* RIR1-1 and *Pfu* RIR1-2 must specifically associate. This indicates that interacting regions of the splicing domains are significantly different between inteins. Specificity may have arisen simply by chance, or it may be critical for the proper functioning of a protein splicing element. For example, it would be detrimental for the splicing domains of two inteins present in the same host protein, such as the *Pfu* RIR1-1 and *Pfu* RIR1-2 inteins, to interact.

## D. Intein-Facilitated Protein Cyclization

Both protein ligation and *trans*-splicing have permitted protein cyclization in vitro and in vivo.[30,50,65,84,85,91−93] Protein ligation requires an N-terminal cysteine and a C-terminal thioester on the same target protein to generate a cyclic species. The N-terminal cysteine can be released by protease treatment or by the action of specially engineered protein splicing elements. The latter technique is utilized in the two-intein or TWIN system.[65] Alternatively, fusion of a split intein to the same target protein can result in a circular protein following *trans*-splicing.[30,50,91,92] This technique was termed SICLOPPS for split intein-mediated circular ligation of peptides and proteins.[50,91]

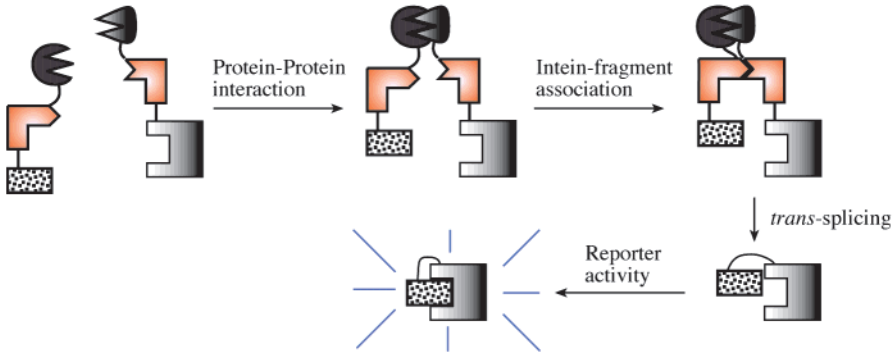


**Figure 7.** Intein-based detection of protein−protein interactions. The test proteins (bait and target) are fused to the appropriate split intein fragments at the gene level. Test protein dimerization facilitates intein splicing domain association. This in turn stimulates splicing and activation of the divided reporter fragments, such as split green fluorescent protein or luciferase.

One goal of these projects was the creation of backbone-constrained peptide libraries. However, most current studies have focused on the effect of backbone cyclization on protein stability. This is an especially relevant question because cyclic peptides have been found in both prokaryotes and eukaryotes. The reason these circular proteins are produced instead of a linear counterpart is not clear. The majority of known naturally occurring cyclic proteins are small, usually around 10−30 amino acid residues, and constraining the ends may allow formation of a stable secondary structure. Biochemical studies found that an intein-cyclized protein, even if it naturally occurs as a linear molecule, is more resistant to heat or denaturants.[85,92] An exception was the cyclization of the N-terminal ca. 60 amino acid residue Src homology domain.[93] One circular form folded properly, but it was not significantly more thermostable than the linear form. Alternatively, another form that lacked a key glutamate residue was considerably more thermostable when cyclized. These studies indicate that in general backbone cyclization increases protein thermostability. The magnitude of this effect may depend on the particular properties of the target protein. Therefore, IPL and *trans*-splicing are novel approaches to cyclize important pharmaceutical peptides or commercial enzymes to increase their stability or potency.

## E. Use of Split Inteins To Screen for Protein Interactions or Contain Transgenes

*Trans*-splicing inteins have been recruited for detecting protein−protein interaction in a two-hybrid system. The yeast two-hybrid system has permitted the determination of a number of new protein−protein interactions.[94] The traditional system is based on triggering the transcription of a reporter, such as $\beta$-galactosidase, by dimerization of two target proteins. In addition to the transcriptional reporter system a number of protein-complementation-based schemes have been developed.[95] These can be used to detect protein−protein interactions in the cytoplasm and at the membrane surface and expand the transcriptional reporter systems. One protein complementation system uses the protein splicing activity of inteins (Figure 7) and either GFP or luciferase as the reporter.[96,97] The reporter proteins are divided into two inactive fragments that upon ligation by *trans*-splicing activate the detectable marker. The test proteins are fused to the appropriate N- or C-terminal splicing domain of the split intein. Interaction of the protein fragments facilitates intein domain association and subsequent *trans*-splicing. Application of this system confirmed the importance of insulin-dependent phosphorylation of Tyr941 in the interaction of the insulin receptor substrate 1 and the SH2 domain of phosphatidylinositol-3-kinase.

A second intein-based methodology merged protein splicing and FKBP12/FRB heterodimerization.[31] The technology used a split *Sce* VMA1 intein. Intein fragment dimerization was accomplished by fusing FKBP and FRB to the N- and C-terminal splicing domains, respectively. In the absence of rapamycin, known to dimerize FKBP and FRB, no *trans*-splicing was detected. However, rapamycin addition triggered the in vitro protein splicing activity of the split *Sce* VMA1 intein. This system expands intein-based methodologies to include screening for protein−protein dimerization induced by small molecules.

A split *Ssp* DnaE intein was also important in a proposed method for preventing transgene spread from a genetically modified plant to weedy relatives. The methodology involves splitting the transgene so that expression of an individual gene fragment results in an inactive truncated protein. Placement of one gene fragment into the nucleus and the complementary fragment into the chloroplast would prevent spread of the complete active gene through pollen.[98,99] For some proteins, activity from the split transgene products may occur by protein complementation. A more general mechanism would be to use protein *trans*-splicing to reconstitute the appropriate activity. Because of the time necessary to perform tests in plants, the initial experiments and proof of principle were performed in *E. coli*. Two commercially relevant transgenes, 5-enolpyruvylshikimate-3-phosphate synthase and acetolactate synthetase, were scanned for appropriate split sites. Sites were found to insert a *trans*-splicing intein and reconstitute activity. Interestingly, *trans*-splicing was not always required for protein activity. This was due to the high affinity of the two intein splicing domains which generate an active protein through intein-

mediated protein complementation. Future studies will test this work in a plant system.

## V. Summary

Much has been learned about inteins since their initial discovery. Finding multiple inteins coupled with mutagenic studies highlighted conserved or functionally important amino acid residues. The relevance of the splice junction amino acid residues became clear with the determination of the protein splicing mechanism. Surprisingly, some inteins were found to use a second mechanism that essentially bypassed the N−S acyl shift important to the majority of protein splicing elements. Crystal structures, kinetic studies, and the discovery of an inhibitor further added to the understanding of inteins and the splicing reaction. Even so, there is much that is still unknown, for example, the role the penultimate histidine or the conserved block B or F residues play in protein splicing. In addition, the potential of inteins as powerful tools for protein engineering was recognized by a number of researchers as indicated by the breadth of intein-based technologies that have been developed. A particularly compelling aspect of inteins is that they can be modified for in vivo and in vitro protein engineering. Therefore, a great deal is left to be discovered, and it will be interesting to see what new intein discoveries or applications will be developed in the future.

## VI. Abbreviations

| | |
|---|---|
| DTT | dithiothreitol |
| MESNA | 2-mercaptoethanesulfonic acid |
| IPL | intein-mediated protein ligation |
| MBP | maltose binding protein |
| SPPL | solid-phase expressed protein ligation |
| TWIN | two-intein |
| SICLOPPS | split intein-mediated circular ligation of peptides and proteins |
| CBD | chitin binding domain |

## VII. Acknowledgments

We thank the past and present members of our laboratories. We also thank Don Comb for his continuing vision and support.

## VIII. References

(1) Perler, F. B.; Davis, E. O.; Dean, G. E.; Gimble, F. S.; Jack, W. E.; Neff, N.; Noren, C. J.; Thorner, J.; Belfort, M. *Nucleic Acids Res.* **1994**, *22*, 1125.
(2) Saldanha, R.; Mohr, G.; Belfort, M.; Lambowitz, A. M. *FASEB J* **1993**, *7*, 15.
(3) Perler, F. B. *Nucleic Acids Res.* **2000**, *28*, 344.
(4) Wu, H.; Hu, Z.; Liu, X. Q. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9226.
(5) Liu, X. Q. *Annu. Rev. Genet.* **2000**, *34*, 61.
(6) Pietrokovski, S. *Trends Genet.* **2001**, *17*, 465.
(7) Derbyshire, V.; Belfort, M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1356.
(8) Perler, F. B. *Cell* **1998**, *92*, 1.
(9) Koufopanou, V.; Goddard, M. R.; Burt, A. *Mol. Biol. Evol.* **2002**, *19*, 239.
(10) Hirata, R.; Ohsumi, Y.; Nakano, A.; Kawasaki, H.; Suzuki, K.; Anraku, Y. *J. Biol. Chem.* **1990**, *265*, 6726.
(11) Kane, P. M.; Yamashiro, C. T.; Wolczyk, D. F.; Neff, N.; Goebl, M.; Stevens, T. H. *Science* **1990**, *250*, 651.
(12) Perler, F. B.; Comb, D. G.; Jack, W. E.; Moran, L. S.; Qiang, B.; Kucera, R. B.; Benner, J.; Slatko, B. E.; Nwankwo, D. O.; Hempstead, S. K.; Carlow, C. K. S.; Jannasch, H. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5577.
(13) Davis, E. O.; Jenner, P. J.; Brooks, P. C.; Colston, M. J.; Sedgwick, S. G. *Cell* **1992**, *71*, 201.
(14) Pietrokovski, S. *Protein Sci.* **1994**, *3*, 2340.
(15) Perler, F. B.; Olsen, G. J.; Adam, E. *Nucleic Acids Res.* **1997**, *25*, 1087.
(16) Pietrokovski, S. *Protein Sci.* **1998**, *7*, 64.
(17) Pietrokovski, S. *Curr. Biol.* **1998**, *8*, R634.
(18) Gorbalenya, A. E. *Nucleic Acids Res.* **1998**, *26*, 1741.
(19) Gimble, F. S.; Thorner, J. *Nature* **1992**, *357*, 301.
(20) Belfort, M.; Roberts, R. J. *Nucleic Acids Res.* **1997**, *25*, 3379.
(21) Shingledecker, K.; Jiang, S.; Paulus, H. *Gene* **1998**, *207*, 187.
(22) Derbyshire, V.; Wood, D. W.; Wu, W.; Dansereau, J. T.; Dalgaard, J. Z.; Belfort, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 11466.
(23) Chong, S.; Xu, M.-Q. *J. Biol. Chem.* **1997**, *272*, 15587.
(24) Telenti, A.; Southworth, M.; Alcaide, F.; Daugelat, S.; Jacobs, J.; William, R.; Perler, F. B. *J. Bacteriol.* **1997**, *179*, 6378.
(25) Smith, D. R.; Doucette-Stamm, L. A.; Deloughery, C.; Lee, H.; Dubois, J.; Aldredge, T.; Bashirzadeh, R.; Blakely, D.; Cook, R.; Gilbert, K.; Harrison, D.; Hoang, L.; Keagle, P.; Lumm, W.; Pothier, B.; Qiu, D.; Spadafora, R.; Vicaire, R.; Wang, Y.; Wierzbowski, J.; Gibson, R.; Jiwani, N.; Caruso, A.; Bush, D.; Safer, H.; Patwell, D.; Prabhakar, S.; McDougall, S.; Shimer, G.; Goyal, A.; Pietrokovski, S.; Church, G. M.; Daniels, C. J.; Mao, J.-I.; Rice, P.; Nolling, J.; Reeve, J. N. *J. Bacteriol.* **1997**, *179*, 7135.
(26) Evans, J., T. C.; Benner, J.; Xu, M.-Q. *J. Biol. Chem.* **1999**, *274*, 3923.
(27) Mills, K. V.; Lew, B. M.; Jiang, S.; Paulus, H. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 3543.
(28) Southworth, M. W.; Adam, E.; Panne, D.; Byer, R.; Kautz, R.; Perler, F. B. *EMBO J.* **1998**, *17*, 918.
(29) Wu, H.; Xu, M.-Q.; Liu, X.-Q. *Biochim. Biophys. Acta* **1998**, *1387*, 422.
(30) Evans, J., T. C.; Martin, D.; Kolly, R.; Panne, D.; Sun, L.; Ghosh, I.; Chen, L.; Benner, J.; Liu, X.-Q.; Xu, M.-Q. *J. Biol. Chem.* **2000**, *275*, 9091.
(31) Mootz, H. D.; Muir, T. W. *J. Am. Chem. Soc.* **2002**, *124*, 9044.
(32) Xu, M.-Q.; Perler, F. B. *EMBO J.* **1996**, *15*, 5146.
(33) Chong, S.; Shao, Y.; Paulus, H.; Benner, J.; Perler, F. B.; Xu, M.-Q. *J. Biol. Chem.* **1996**, *271*, 22159.
(34) Southworth, M. W.; Benner, J.; Perler, F. B. *EMBO J.* **2000**, *19*, 5019.
(35) Yamamoto, K.; Low, B.; Rutherford, S. A.; Rajagopalan, M.; Madiraju, M. V. *Biochem. Biophys. Res. Commun.* **2001**, *280*, 898.
(36) Hodges, R. A.; Perler, F. B.; Noren, C. J.; Jack, W. E. *Nucleic Acids Res.* **1992**, *20*, 6153.
(37) Hirata, R.; Anraku, Y. *Biochem. Biophys. Res. Commun.* **1992**, *188*, 40.
(38) Xu, M.-Q.; Comb, D. G.; Paulus, H.; Noren, C. J.; Shao, Y.; Perler, F. B. *EMBO J.* **1994**, *13*, 5517.
(39) Xu, M. Q.; Southworth, M. W.; Mersha, F. B.; Hornstra, L. J.; Perler, F. B. *Cell* **1993**, *75*, 1371.
(40) Cooper, A. A.; Chen, Y. J.; Lindorfer, M. A.; Stevens, T. H. *EMBO J.* **1993**, *12*, 2575.
(41) Wallace, C. J. *Protein Sci.* **1993**, *2*, 697.
(42) Shao, Y.; Xu, M.-Q.; Paulus, H. *Biochemistry* **1995**, *34*, 10844.
(43) Klabunde, T.; Sharma, S.; Telenti, A.; Jacobs, W. R. J.; Sacchettini, J. C. *Nat. Struct. Biol.* **1998**, *5*, 31.
(44) Duan, X.; Gimble, F. S.; Quiocho, F. A. *Cell* **1997**, *89*, 555.
(45) Poland, B. W.; Xu, M.-Q.; Quiocho, F. A. *J. Biol. Chem.* **2000**, *275*, 16408.
(46) Mizutani, R.; Nogami, S.; Kawasaki, M.; Ohya, Y.; Anraku, Y.; Satow, Y. *J. Mol. Biol.* **2002**, *316*, 919.
(47) Chong, S.; Williams, K. S.; Wotkowicz, C.; Xu, M.-Q. *J. Biol. Chem.* **1998**, *273*, 10567.
(48) Chen, L.; Benner, J.; Perler, F. B. *J. Biol. Chem.* **2000**, *275*, 20431.
(49) Wang, S.; Liu, X. Q. *J. Biol. Chem.* **1997**, *272*, 11869.
(50) Scott, C. P.; Abel-Santos, E.; Wall, M.; Wahnon, D. C.; Benkovic, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13638.
(51) Mills, K. V.; Paulus, H. *J. Biol. Chem.* **2001**, *276*, 10832.
(52) Kawasaki, M.; Nogami, S.; Satow, Y.; Ohya, Y.; Anraku, Y. *J. Biol. Chem.* **1997**, *272*, 15668.
(53) Ghosh, I.; Sun, L.; Xu, M. Q. *J. Biol. Chem.* **2001**, *276*, 24051.
(54) Carter, P.; Wells, J. A. *Nature* **1988**, *332*, 564.
(55) Xu, Q.; Buckley, D.; Guan, C.; Guo, H. C. *Cell* **1999**, *98*, 651.
(56) Outten, C. E.; O'Halloran, T. V. *Science* **2001**, *292*, 2488.
(57) Zubay, G. *Biochemistry*, 2nd ed.; Macmillan Publishing Co.: New York, 1988.
(58) Paulus, H. *Bioorg. Chem.* **2001**, *29*, 119.
(59) Shingledecker, K.; Jiang, S.; Paulus, H. *Arch. Biochem. Biophys.* **2000**, *375*, 138.
(60) Martin, D. D.; Xu, M.-Q.; Evans, T. C., Jr. *Biochemistry* **2001**, *40*, 1393.
(61) Southworth, M. W.; Amaya, K.; Evans, J., T. C.; Xu, M.-Q.; Perler, F. B. *BioTechniques* **1999**, *27*, 110.

(62) Mathys, S.; Evans, J., T. C.; Chute, I.; Wu, H.; Chong, S.; Benner, J.; Liu, X.-Q.; Xu, M.-Q. *Gene* **1999**, *231*, 1.
(63) Wood, D. W.; Wu, W.; Belfort, G.; Derbyshire, V.; Belfort, M. *Nat. Biotechnol.* **1999**, *17*, 854.
(64) Mercure, S.; Lafontaine, D.; Ananvoranich, S.; Perreault, J. P. *Biochemistry* **1998**, *37*, 16975.
(65) Evans, T. C.; Benner, J.; Xu, M.-Q. *J. Biol. Chem.* **1999**, *274*, 18359.
(66) 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.; Pelletier, J. J.; Paulus, H.; Xu, M.-Q. *Gene* **1997**, *192*, 271.
(67) Watanabe, T.; Ito, Y.; Yamada, T.; Hashimoto, M.; Sekine, S.; Tanaka, H. *J. Bacteriol.* **1994**, *176*, 4465.
(68) Dubendorff, J. W.; Studier, F. W. *J. Mol. Biol.* **1991**, *219*, 45.
(69) Chong, S.; Montello, G. E.; Zhang, A.; Cantor, E. J.; Liao, W.; Xu, M.-Q.; Benner, J. *Nucleic Acids Res.* **1998**, *26*, 5109.
(70) Zhang, A.; Gonzalez, S. M.; Cantor, E. J.; Chong, S. *Gene* **2001**, *275*, 241.
(71) Wood, D. W.; Derbyshire, V.; Wu, W.; Chartrain, M.; Belfort, M.; Belfort, G. *Biotechnol. Prog.* **2000**, *16*, 1055.
(72) Dawson, P. E.; Muir, T. W.; Clark-Lewis, I.; Kent, S. B. *Science* **1994**, *266*, 776.
(73) Tam, J. P.; Lu, Y.-A.; Liu, C.-F.; Shao, J. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 12485.
(74) Evans, J., T. C.; Benner, J.; Xu, M.-Q. *Protein Sci.* **1998**, *7*, 2256.
(75) Muir, T. W.; Sondhi, D.; Cole, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 6705.
(76) Severinov, K.; Muir, T. W. *J. Biol. Chem.* **1998**, *273*, 16205.
(77) Cottingham, I. R.; Millar, A.; Emslie, E.; Colman, A.; Schnieke, A. E.; McKee, C. *Nat. Biotechnol.* **2001**, *19*, 974.
(78) Zhu, H.; Bilgin, M.; Bangham, R.; Hall, D.; Casamayor, A.; Bertone, P.; Lan, N.; Jansen, R.; Bidlingmaier, S.; Houfek, T.; Mitchell, T.; Miller, P.; Dean, R. A.; Gerstein, M.; Snyder, M. *Science* **2001**, *293*, 2101.
(79) Lesaicherre, M.-L.; Lue, R. Y. P.; Chen, G. Y. J.; Zhu, Q.; Yao, S. Q. *J. Am. Chem. Soc.* **2002**, *124*, 8768.
(80) Sydor, J. R.; Mariano, M.; Sideris, S.; Nock, S. *Bioconjugate Chem.* **2002**, *13*, 707.
(81) Cotton, G. J.; Ayers, B.; Xu, R.; Muir, T. W. *J. Am. Chem. Soc.* **1999**, *121*, 1100.
(82) Erlanson, D. A.; Chytil, M.; Verdine, G. L. *Chem. Biol.* **1996**, *3*, 981.
(83) Xu, R.; Ayers, B.; Cowburn, D.; Muir, T. W. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 388.
(84) Camarero, J. A.; Muir, T. W. *J. Am. Chem. Soc.* **1999**, *121*, 5597.
(85) Iwai, H.; Plueckthun, A. *FEBS Lett.* **1999**, *459*, 166.
(86) Cotton, G. J.; Muir, T. W. *Chem. Biol.* **2000**, *7*, 253.
(87) Lew, B. M.; Mills, K. V.; Paulus, H. *J. Biol. Chem.* **1998**, *273*, 15887.
(88) Otomo, T.; Teruya, K.; Uegaki, K.; Yamazaki, T.; Kyogoku, Y. *J. Biomol. NMR* **1999**, *14*, 105.
(89) Yamazaki, T.; Otomo, T.; Oda, N.; Kyogoku, Y.; Uegaki, K.; Ito, N.; Ishino, Y.; Nakamura, H. *J. Am. Chem. Soc.* **1998**, *120*, 5591.
(90) Otomo, T.; Ito, N.; Kyogoku, Y.; Yamazaki, T. *Biochemistry* **1999**, *38*, 16040.
(91) Scott, C. P.; Abel-Santos, E.; Jones, A. D.; Benkovic, S. J. *Chem. Biol.* **2001**, *8*, 801.
(92) Williams, N. K.; Prosselkov, P.; Liepinsh, E.; Line, I.; Sharipo, A.; Littler, D. R.; Curmi, P. M.; Otting, G.; Dixon, N. E. *J. Biol. Chem.* **2002**, *277*, 7790.
(93) Camarero, J. A.; Fushman, D.; Cowburn, D.; Muir, T. W. *Bioorg. Med. Chem.* **2001**, *9*, 2479.
(94) Fields, S.; Song, O. *Nature* **1989**, *340*, 245.
(95) Pelletier, J. N.; Campbell-Valois, F. X.; Michnick, S. W. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 12141.
(96) Ozawa, T.; Takeuchi, T. M.; Kaihara, A.; Sato, M.; Y., U. *Anal. Chem.* **2001**, *73*, 5866.
(97) Ozawa, T.; Kaihara, A.; Sato, M.; Tachihara, K.; Umezawa, Y. *Anal. Chem.* **2001**, *73*, 2516.
(98) Sun, L.; Ghosh, I.; Paulus, H.; Xu, M. Q. *Appl. Environ. Microbiol.* **2001**, *67*, 1025.
(99) Chen, L.; Pradhan, S.; Evans, T. C. J. *Gene* **2001**, *263*, 39.

CR9601369