# PROTEIN SPLICING: MECHANISM AND APPLICATIONS IN BIOORGANIC CHEMISTRY

Reported by Kimberly L. Berkowski   February 19, 2001

INTRODUCTION

The discovery of protein splicing has altered the central dogma of gene expression.[1] According to this dogma, genetic information flows from DNA to RNA through transcription, and is then translated and expressed as protein. However, more genetic information is typically coded for than actually appears in the protein product. This information is posttranscriptionally excised as introns, in a process termed RNA splicing. Just over ten years ago, two groups independently discovered that some proteins are capable of excising genetic information *posttranslationally* through a process analogous to RNA splicing.[2,3] In protein splicing internal segments in the posttranslational product (inteins) excise themselves from the protein while ligating the flanking polypeptides (exteins) to form the final protein product (Figure 1). This excision and ligation process is novel because it is *self-catalyzed*.[4] The discovery of protein splicing and elucidation of its mechanism triggered the research of new applications and techniques for protein chemistry and engineering.

Figure 1. Protein Biosynthesis.
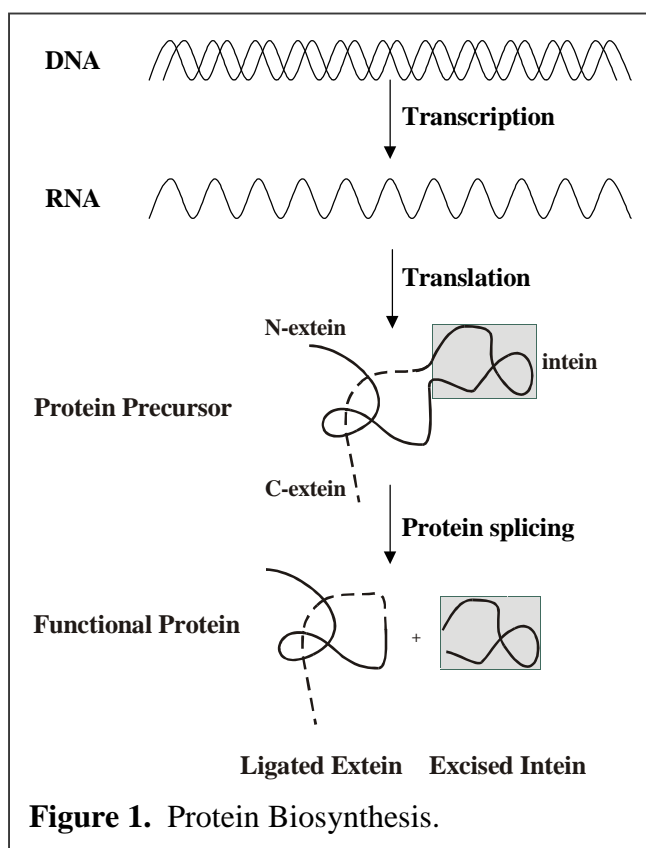
INTEINS: OCCURRENCE AND STRUCTURE

To date, over 100 inteins have been discovered - all from proteins of unicellular organisms.[5] Approximately 70% of inteins reside in host proteins involved in DNA replication and repair. The reasons for this intein preference are unclear, although many hypotheses have been proposed.[6] Inteins have been found to contain 134-608 amino acids. They are usually composed of two distinct regions: a protein splicing domain (which is split into two segments) and an endonuclease domain.[5] Smaller inteins, however, contain only the split protein splicing domain separated by a short amino acid linker

(Figure 2).[7] The protein splicing domain catalyzes its own excision from the protein as well as the ligation of the flanking exteins. The endonuclease region is a mobile genetic element that duplicates itself and inserts into a homologous, but intein-lacking allele.[6] The two domains of the intein function independently of each other.[7] This review will focus only on the protein splicing region of the intein.

Figure 2. Intein Structure.
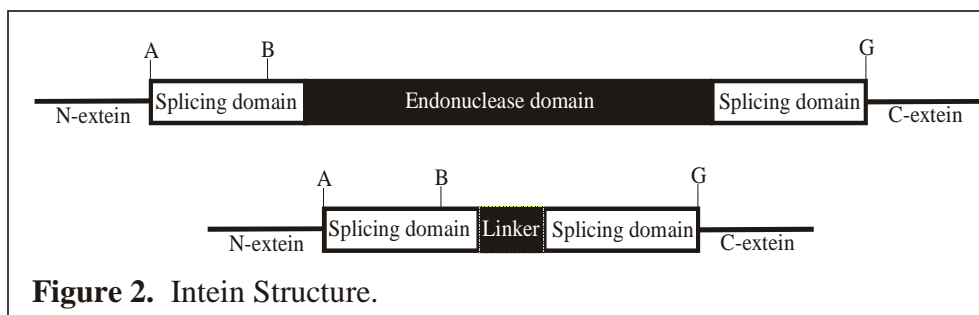
Certain amino acids are highly conserved within the protein splicing domain. The first amino acid on the N-terminal intein (block A) and the C-terminal extein is either serine, cysteine, or threonine, although an alanine residue has sometimes been observed at the N-terminal splice junction.[1] The amino acid on the C-terminal of the intein (block G) is most commonly asparagine, preceded by a histidine.[7] Finally, a Thr-x-x-His sequence in block B of the N-terminal intein is usually observed (Figure 2).[1] These residues play specific roles in the structure of the intein and the mechanism of protein splicing as discussed below.

MECHANISM

Protein splicing occurs via four independent,[4] intramolecular[8] reactions. The first three reactions are catalyzed by the intein at a single site.[7] The final reaction is a spontaneous rearrangement to form a native peptide bond. Although the chemical reactions involved in protein splicing are well understood, the mechanism of catalysis for each reaction is just now beginning to be elucidated.
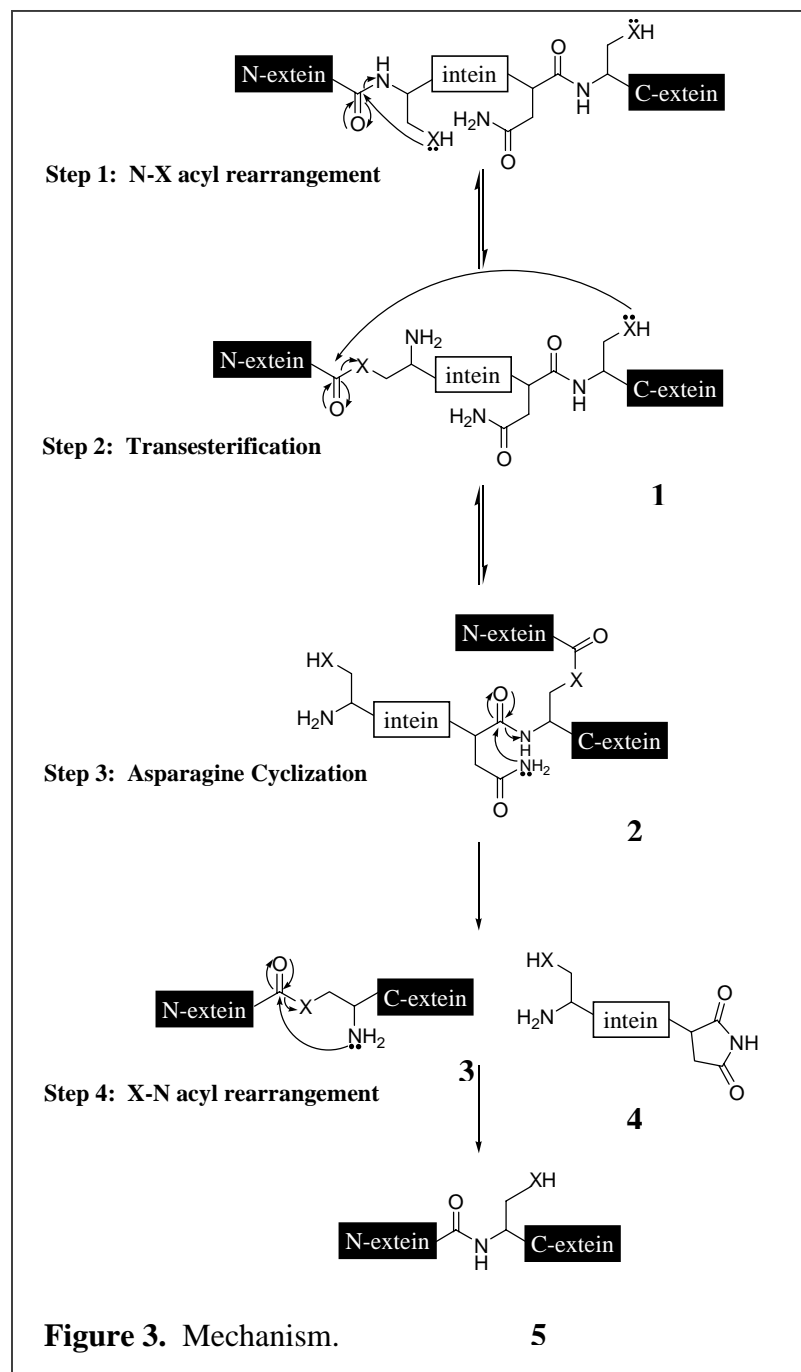
N → X Acyl Rearrangement

The first step of protein splicing involves the nucleophilic attack of the first intein residue, the conserved cys/ser amino acid, on the adjacent carbonyl carbon of the extein (Figure 3). The reaction goes through an oxyoxazolidine or oxythiazolidine anion to form **1**, an ester or thioester.[1] Because the equilibrium constant is typically unfavorable for the rearrangement of an amide to an ester or thioester, the peptide bond at the N-terminal splice junction has been proposed to exist in a high energy conformation.[7]

The crystal structures of a mutant GyrA protein from *Mycobacterium xenopi* and a mutant VMA protein from *Saccharomyces cerevisiae* (VMA29) support this hypothesis. The crystallized GyrA intein

has one N-terminal extein residue, 197 intein residues, and no C-terminal extein residues. In addition, the first intein residue was mutated from cysteine to serine to retard splicing.[9] In the crystallized VMA29 protein, the cysteine residue at the N-terminal splice junction, and the asparagine residue at the C-terminal splice junction were mutated to alanine to prevent splicing. The N-terminal extein contains 10 amino acid residues and the C-termial extein contains four amino acid residues. The crystal structure of the GyrA intein shows the scissile peptide bond at the N-terminal splice junction in a less energetically favorable *cis* conformation.[9] This bond conformation correctly positions the serine residue at the N-terminal splice junction for nucleophilic attack on the adjacent carbonyl carbon of the extein.[9] The crystal structure of VMA29 however, indicates that a distorted *trans* conformation exists at

Figure 3. Mechanism.

the N-terminal splice junction.[10] These crystal structures support the hypothesis that ester formation may occur to relieve the strain of a high energy peptide bond.

The crystal structures of the GyrA and VMA29 inteins also help to explain the unfavorable amide-ester equilibrium shift by showing amino acid residues that could potentially stabilize the tetrahedral intermediate through hydrogen bonding via an oxyanion hole.[10] Stabilization of the tetrahedral intermediate lowers the activation energy of the reaction and facilitates ester formation. Finally, both crystal structures show a conserved histidine residue positioned within hydrogen bonding distance to the nitrogen of the scissile peptide bond. This histidine residue may catalyze the reaction by

donating a proton to the amine, making it a better leaving group. Although much insight has been recently obtained for the catalysis of the N → X acyl rearrangement, the mechanism remains somewhat ambiguous.

Transesterification

In the second step of protein splicing, the conserved cys/ser/thr amino acid on the C-terminal extein of **1** attacks the newly formed ester group of the N-terminal splice junction. This transformation results in the ligation of the exteins as shown by **2**. The crystal structure of VMA29 shows that a 9 Å gap exists between the nucleophile and electrophile.[10] How this gap is transversed by the nucleophilic residues is unclear. However, Poland and co-workers propose that a conformational change after the first step brings the splice junctions into closer proximity. The crystal structure of the VMA29 intein shows the presence of a zinc ion at the C-terminal splice junction. The zinc ion is coordinated to the nucleophilic cysteine (Cys^{455}) involved in transesterification, the penultimate histidine residue (His^{453}), a glutamic acid (Glu^{80}), and a water molecule. This coordination strains the backbone of the protein and may facilitate transesterification.[10] In addition, Shingledecker and coworkers determined that the pK_a of the sulfhydryl group on Cys^{455} is 5.8. Typically, the sulfhydryl groups in proteins have pK_a values of 8. Thus, this result correlates well with the discovery of the coordination of cysteine to zinc.

Asparagine Cyclization

In the third step of protein splicing the amide side chain of asparagine at the C-terminal splice junction attacks its own carbonyl carbon on the main chain. The asparagine residue cyclizes, forming a C-teminal aminosuccinimide, **3**, and the excised intein **4**.[8]

The crystal structure of the VMA29 intein has provided insight to the mechanism and catalysis of asparagine cyclization. Poland and coworkers suggest that transesterification to form the branched intermediate disrupts the coordination between the cysteine residue and the zinc ion. As a result zinc may move, coordinate to His^{442} and one of amino acid residues flanking Cys^{455}. As a result, the penultimate histidine (His^{453}) may move to a position where it is able to protonate the main chain amide of Asn^{454} after cyclization, catalyzing peptide hydrolysis. In addition, the reorientation of zinc as well as some hydrophobic interactions may correctly position the β-amide of asparagine to attack its backbone carbonyl carbon.[10]

X → N Acyl Rearrangement

In the final step of protein splicing the ester moiety rearranges to form a protein with a more thermodynamically stable amide group, **5**. Unlike the first step of protein splicing, an intein is not present to stabilize the ester group. Therefore, this step is spontaneous and virtually irreversible.[1]

APPLICATIONS

The discovery that inteins self-catalyze their excision from proteins has led to the development of interesting and useful applications in protein science and engineering. Two general types of applications stem from protein splicing. The first application involves the mutation of either the N- or C-terminal splice junction to inhibit protein splicing. Controlled cleavage is still allowed, however, at the unmutated splice junction. In the second application, protein splicing is induced by the reassembly of a split intein.

Intein-Mediated Affinity Chromatography for Protein Purification

Intein-mediated protein purification is a common application of protein splicing. With this technique, specific reagents, temperature, or pH are used to controllably cleave the intein at either its N- or C-terminal splice junction.[13,14] Intein-mediated protein purification systems work by fusing one terminus of the intein to an affinity tag and the other terminus to a target protein. After adsorption onto an affinity column, a thiol reagent is introduced. The thiol reagent attacks the thioester moiety resulting from N $\rightarrow$ S acyl rearrangement and cleaves the protein off the affinity column. Finally, the purified protein is released from the thiol reagent by hydrolysis. Intein-mediated affinity columns are advantageous because, unlike traditional protein purification systems, they eliminate the need for proteases, which can be unspecific and result in undesirable products. In addition, because inteins are self-catalytic, protein purification can occur in one step.[13,14]

Intein-Mediated Protein Ligation

A major breakthrough in protein synthesis occurred with the development of native chemical ligation, wherein a synthetic peptide with a C-terminal α-thioester is reacted with a peptide containing an N-terminal cysteine residue (Figure 4). The resulting thioester-linked intermediate undergoes spontaneous rearrangement affording a native peptide bond.[15] Native chemical ligation is limited because peptides with the α-thioester moieties must be chemically synthesized.

The ability of inteins to generate α-thioesters when immobilized on affinity columns was used to address this problem (Figure 5). In this method, the target protein is fused to the N-terminus of the intein and the C-terminus of the intein is linked to an affinity tag. The protein is then purified on an affinity column and released from the intein by the introduction of a thiol-cleaving reagent, leaving a peptide with a C-terminal α-thioester. At this stage, a peptide with an N-terminal cysteine is introduced to the system and its thiol side chain attacks the α-thioester functionality. The resulting linked thioester intermediate spontaneously rearranges to form a native peptide bond.[16,17]

Figure 4.
Native Chemical Ligation

**Figure 5.** Intein-Mediated Protein Ligation.
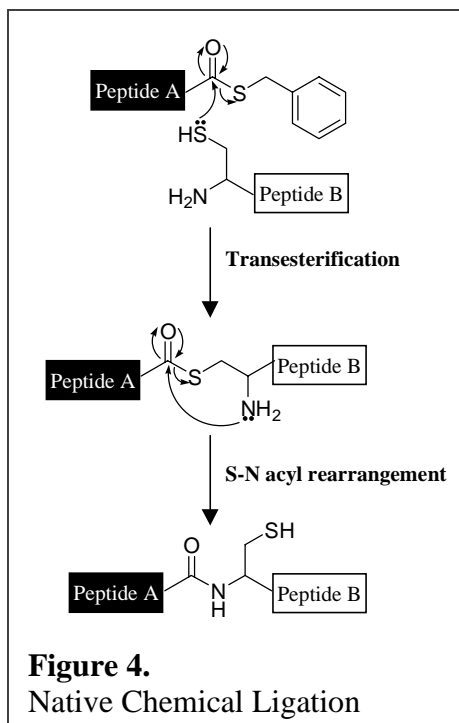
Peptides with N-terminal cysteine residues were generated in a similar manner (Figure 5). In this case the target protein is linked to the C-terminus of the intein and the affinity tag is linked to the mutated N-terminus of the intein. The protein is then purified on an affinity column and asparagine cyclization releases a peptide with an N-terminal cysteine residue.[18]

Intein-mediated protein ligation is an extraordinarily useful method for protein synthesis with a variety of peripheral applications. Some of these applications include protein cyclization via a two intein (TWIN) system,[19] synthesis of multi-domain proteins,[20] and isotopic labeling of protein segments for NMR spectroscopy.[21]

Protein *trans*-Splicing

Paulus and coworkers have found that protein splicing can be triggered by reassociation of two inactive segments of a split intein.[22] "Protein *trans*-splicing" is a method used to ligate any two unrelated peptides. In *trans*-splicing, the intein of the precursor protein is split into two segments. Each segment is fused to an appropriate extein and expressed in a different host. Upon recombination of the denatured intein fragments, the fragments reassociate and are able to initiate protein splicing without being covalently linked (Figure 6).[23] Paulus and coworkers found that if renaturation is carried out in
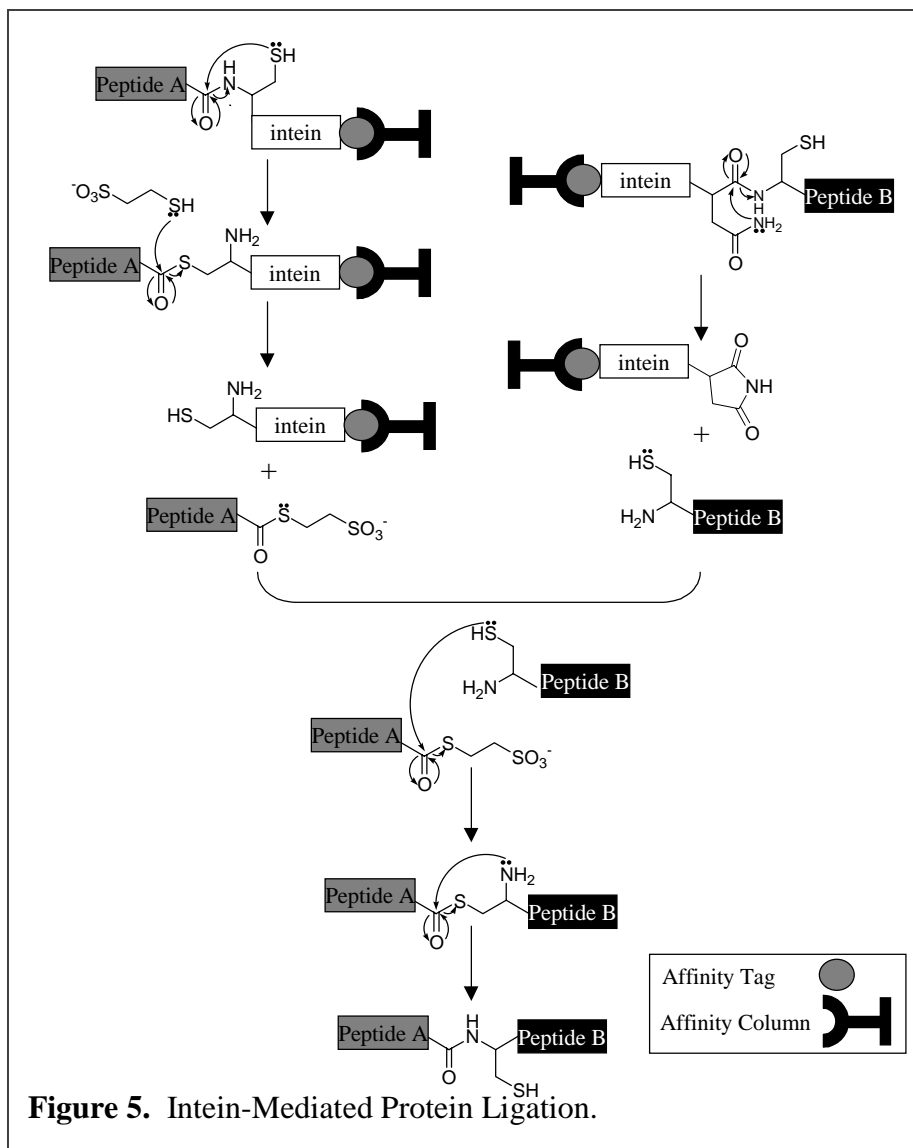
the absence of a reductant, a disulfide-linked dimer at the splice junction of the two intein fragments results.[24] Protein splicing is then inhibited until the disulfide moiety is reduced. Although protein *trans*-splicing was first demonstrated *in vitro*, an example of a naturally split intein that performed *trans*-splicing has also been discovered in a DnaE protein of the cyanobacterium *Synechocystis*.[25]

Figure 6. *Trans*-Splicing.
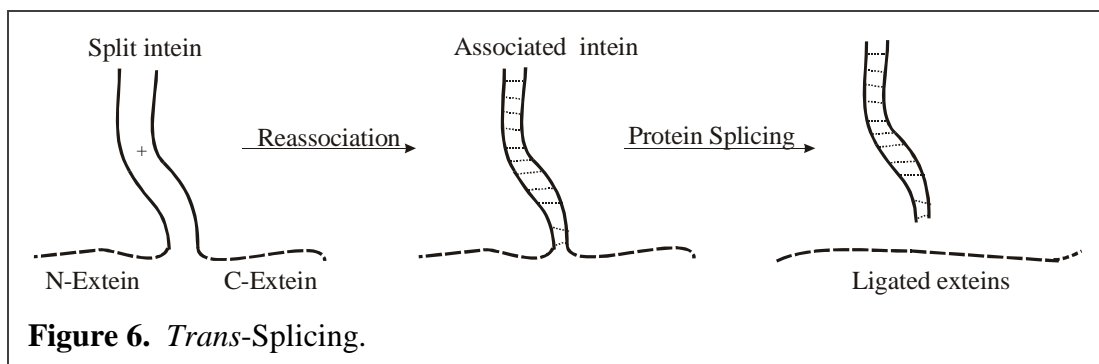
Protein *trans*-splicing has been extended to other techniques in protein engineering. Split inteins may be used to generate cyclic peptides in a process called *in vivo* split intein-mediated circular ligation of peptides and proteins, or SICLOPPS.[26] Split inteins may also be used to segmentally label proteins for NMR spectroscopy.[27] Most recently, *trans*-splicing has been used as a method for monitoring protein-protein interactions *in vivo*.[28] The inteins of these systems are conjugated to biomolecules that specifically interact with each. The exteins of these systems are two segments of a protein that fluoresce upon ligation. The specific interaction of the biomolecules conjugated on the inteins initiates assembly of the intein fragments, which induces protein splicing. As a result, the exteins are ligated and the molecule fluoresces.[28]

CONCLUSIONS

Through the discovery of protein splicing and the elucidation of its mechanism, chemists received a glimpse of how nature engineers proteins. Over the past ten years chemists have used this knowledge to advance their own methods of protein synthesis. This new understanding has allowed the development of new techniques for protein manipulation, enabling them to extend protein synthesis into novel applications.

REFERENCES

(1)   Noren, C.J.; Wang, J.; Perler, F.B. *Angew. Chem. Int. Ed*. **2000**, *39*, 450-466.
(2)   Kane, P. M.; Yamashiro, C. T.; Wolczyk, D. F.; Neff, N.; Goebl, M.; Stevens, T. H. *Science* **1990**, *250*, 651-657.
(3)   Hirata, R.; Ohsumi, Y.; Nakano, A.; Kawasaki, H.; Suzuki, K.; Anraku, Y. *J. Biol. Chem*. **1990**,

265, 6726-6733.

(4)   Xu, M.-Q.; Southworth, M. W.; Mersha, F. B.; Hornstra, L. J.; Perler, F. B. *Cell* **1993**, *75*, 1371-1377.

(5)   InBase, The New England Biolabs Intein Database. http://www.neb.com/neb/inteins.html

(6)   Liu, X. Q. *Annu. Rev. Genet*. **2000**, *34*, 61-76.

(7)   Paulus, H. *Annu. Rev. Biochem*. **2000**, *69*, 447-496.

(8)   Kawaski, M.; Satow, Y.; Ohya, Y.; Anraku, Y. *FEBS Lett*. **1997**, *412*, 518-520.

(9)   Klabunde, T.; Sharma, S.; Telenti, A.; Jacobs Jr., W. R.; Sacchettini, J. C. *Nat. Struct. Biol.* **1998**, *5*, 31-36.

(10)   Poland, B. W.; Xu, M-Q.; Quiocho, F. A. *J. Biol. Chem*. **2000**, *275*, 16408-16413.

(11)   Shingledecker, K.; Jiang, S-Q.; Paulus, H. *Arch. Biochem. Biophys*. **2000**, *375*, 138-144.

(12)   Chen, L.; Benner, J.; Perler, F.B. *J. Biol. Chem*. **2000**, *275*, 20431-20435.

(13)   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*, 271-281.

(14)   Chong, S.; Montello, G. E.; Zhang, A.; Cantor, E. J.; Liao, W.; Xu, M-Q.; Benner, J. *Nucleic Acids Res*. **1998**, *26*, 5109-5115.

(15)   Dawson, P. E.; Muir, T. W.; Clark-Lewis, I.; Kent, S. B. G. *Science* **1994**, *266*, 776-779.

(16)   Muir, T. W.; Sondhi, D.; Cole, P. A. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6705-6710.

(17)   Evans, Jr., T. C.; Benner, J.; Xu, M-Q. *Protein Sci*. **1998**, *7*, 2256-2264.

(18)   Mathys, S.; Evans, Jr.; T. C.; Chute, I. C.; Wu, H.; Chong, S.; Benner, J.; Liu, X-Q.; Xu, M-Q. *Gene*, **1999**, *231*, 1-13.

(19)   Evans, Jr., T. C.; Xu, M-Q. *Biopolymers* **1999**, *51*, 333-342.

(20)   Blaschke, U. K.; Cotton, G. J. Muir, T. W. *Tetrahedron*, **2000**, *56*, 9461-9470.

(21)   Otomo, T.; Ito, N.; Kyogoku, Y.; Yamazaki, T. *Biochemistry* **1999**, *38*, 16040-16044.

(22)   Shingledecker, K.; Jiang, S-Q.; Paulus, H. *Gene* **1998**, *207*, 187-195.

(23)   Lew, B. M.; Mills, K. V.; Paulus, H. *Biopolymers* **1999**, *51*, 355-362.

(24)   Mills, K. V.; Lew, B. M.; Jiang, S-Q.; Paulus, H. *Proc. Natl. Acad. USA* **1998**, *95*, 3543-3548.
        Lew, B. M.; Mills, K. V.; Paulus, H. *J. Biol. Chem*. **1998**, *273*, 15887-15890.

(25)   Perler, F. B. *Trends Biochem. Sci*. **1999**, *24*, 209-211.

(26)   Scott, C. P.; Abel-Santos, E.; Wall, M.; Wahnon, D. C.; Benkovic, S. J. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 13638-13643.

(27)   Yamazaki, T.; Otomo, T.; Oda, N.; Kyogoku, Y.; Uegaki, K.; Ito, N.; Ishino, Y.; Nakamura, H. *J. Am. Chem. Soc*. **1998**, *120*, 5591-5592.

(28)   Ozawa, T.; Nogami, S.; Sato, M.; Ohya, Y.; Umezawa, Y. *Anal. Chem*. **2000**, *72*, 5151-5157.