MJS, Vol. 3, No. 1, May. 2004 - Oct. 2004 pp. 94-106
ISSN 0975-3311
https://doi.org/10.12725/mjs.5.11

PROTEIN SPLICING

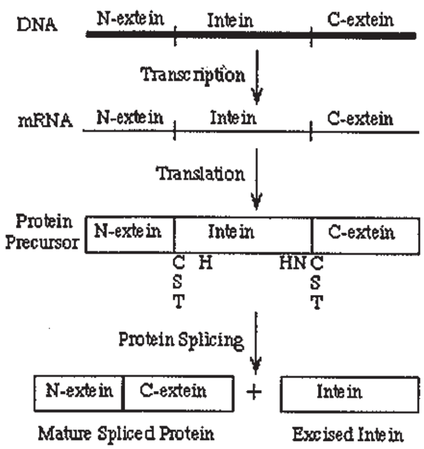
* ND Deve Gowda

Protein splicing is defined as the excision of an intervening sequence (intein) from a protein precursor and the concomitant ligation of the flanking protein fragments (extein) to form a mature host protein (extein) and the free intein (Perler 1994). Intein-mediated protein splicing results in a native peptide bond between the ligated exteins (Cooper 1993). Extein ligation differentiates protein splicing from other forms of autoproteolysis and conserved intein motifs differentiate inteins from other types of in-frame sequences present in one homolog and absent in another homolog.

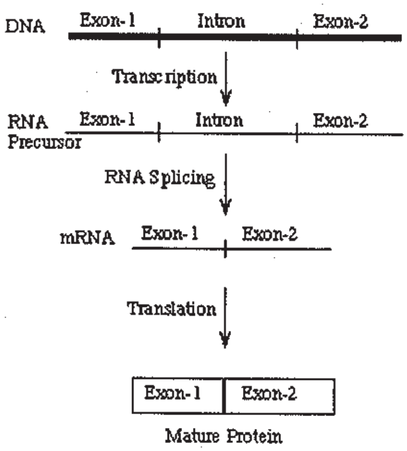
When we compare the RNA splicing and protein splicing as depicted in the following illustration, the introns are intervening sequences that are spliced out of RNA before the mRNA is translated into a protein. The intron and the exon usually do not form a single open reading frame (ORF). During intein-mediated protein splicing, the intervening sequence is both present in the mature mRNA and translated to form a precursor protein. The intein is then spliced out of the precursor protein. The intein plus the first C-extein residue (called the +1 amino acid) contain sufficient information to mediate splicing of the intein out of the host protein and ligation of the exteins to form the active host protein.

* Lecturer in Genetics, Department of Life Sciences, Sri Bhagawan Mahaveer Jain College, Dr. ANK Road, VV Puram, Bangalore 4

Protein Splicing:



RNA Splicing:



Features and roles of inteins

Inteins are selfish genetic elements. They code for proteins that catalyze their excision out of the host proteins,ligating the host flanks with a polypeptide bond. This protein splicing activity is autoproteolytic and is not dependent on any host specific factors. Inteins are very diverse in sequence but all have a protein splicing activity i.e. simple to assay. More than 140 inteins are known from bacteria, archaea and lower eukaryotes. Inteins protein splicing are excellent system for studying protein sequence or functional relation. Inteins are known to naturally protein splice in the cytoplasm of multicellular organisms. Inteins are partial sequences.

The precursor and intermediates or slide products of the reaction corresponding to N-terminal and C-terminal intein cleavage without ligation were also seen. The splicing is required to preserve the functional integrity of the host protein. Endonucloses can be viewed as a substantial driving force in molecular evolution. Through this capacity to make nicks and breaks in DNA, endonuclese genes can invade sequences to form molecular associations that not only mobilize introns and inteins but can also provide catalytic function to other proteins. e.g. HNH endonuclease cassette 'colicin family'.

Intron endonuclease can provide selective advantages in both phage and archael systems whereas colicins promote host defense thereby influencing the stability.

Propensity of endonuclease genes to colonize genomes, can influence genome stability and configurations by promoting lesions in DNA and subsequent intron-intein based arrangements.

Mini inteins that are TS+ indicate that splicing function is contained within the first 94 and last 35aa of the intein. Derivatives of the products of the splicing reaction, ligated exteins and free intein were readily detected on coomassie gels and their identity was verified by western blot analysis

Nomenclature of inteins

Inteins are named after the organism and gene in which they are found. The organism name follows the same consensus as restriction enzymes and uses a 3 letter genus + species designation, followed by a strain designation, if necessary. The organism name is followed by an abbreviation of the extein name. If more than 1 intein is present in an extein gene, the inteins are given a numerical suffix starting from 5' to 3' or in order of their identification.

For example, the Pyrococcus furiosus ribonucleoside-diphosphate reductase alpha subunit gene contains 2 inteins. The organism is abbreviated as 'Pfu'. Since the gene has been called the 'RIR1' gene, the inteins are named using this gene name. Thus, these 2 inteins are called the Pfu RIR1-1 intein inserted after Gly 301 in the Pfu RIR1 precursor protein and the Pfu RIR1-2 intein inserted after Pro914 in the Pfu RIR1 precursor protein.

Note that an intein name, such as the Pfu RIR1-1 intein, refers to both the intein gene and the intein protein. In many publications, the consensus is to italicize the gene name and to capitalize the first letter of the protein name.

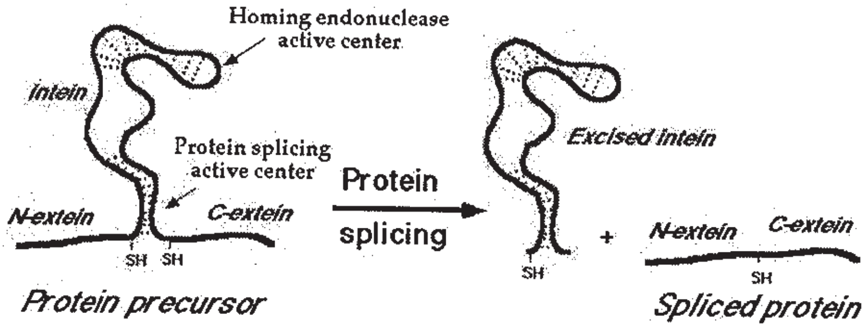
As described below, some inteins are bifunctional proteins that also have endonuclease activity. When endonuclease activity has been demonstrated, the intein is also given a second name that follows the endonuclease naming conventions (Belfort 1997). This name includes the prefix 'PI-', the 3 letter organism abbreviation and a Roman numeral indicating the order of identification of the intein endonuclease in that organism. The endonuclease names for the Pfu RIR1- and Pfu RIR1-2 inteins are PI-PfuI and PI-PfuII, respectively.

There is also a convention for numbering amino acids in inteins. Although we often number the residues in the precursor as a single protein, as when intein insertion site locations are given, a second numbering scheme is often used to assist thinking about inteins in heterologous or foreign exteins. The intein amino acids are numbered from N-terminal to C-terminal beginning with the first residue of the intein and ending with the last residue of the intein. The amino acids in the N-extein: (a) start with the number 1, (b) include a minus sign prefix and (c) are counted from right to left (beginning with the last N-extein residue and going towards the N-terminus). The amino acid preceding the intein is the -1 amino acid. The amino acids in the C-extein: (a) are numbered beginning at the C-terminal splice junction, (b) include a plus sign prefix and (c) are counted from amino to C-terminus.

The first residue following the intein is the mechanistically essential +1 amino acid, which is not technically part of the intein since the intein is defined as the intervening sequence that is spliced out of the precursor.

Mechanism

General mechanism



It involves the precise, self-catalyzed excision of an intervening polypeptide sequence, the intein, from an inactive precursor protein with the concomitant joining of the flanking sequences, the exteins, to produce a new functional protein. All information and catalytic groups required for protein splicing reside in the intein and the two flanking amino acids. The excised intein often functions as homing endonuclease, a property which makes inteins infectious elements that can be transferred horizontally between organisms and even species.

All information and catalytic groups required for protein splicing reside in the intein. In the period 1993-96, we succeeded in defining each of the steps in the protein splicing process by isolating and characterizing the reaction intermediates. Protein splicing is a complex four-step process, which involves;

1. N-S or N-O rearrangement of a peptide bond adjacent to a Cys or Ser residue to yield a linear peptide ester,

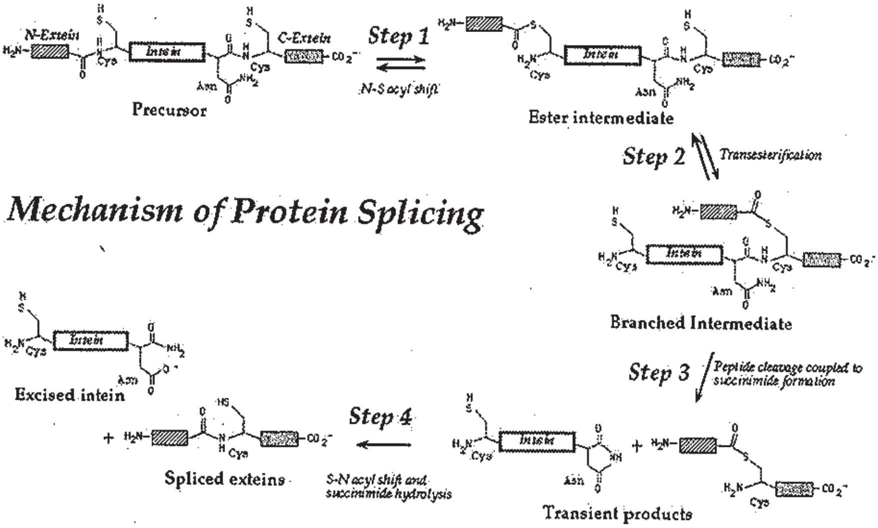
2. Transesterification with a Cys, Ser, or Thr residue at the downstream splice junction to yield a branched ester intermediate,

3. Cyclization of an Asn residue coupled to peptide bond cleavage, and

4. Rearrangement of the transient splicing products to yield stable polypeptides. The first three reactions are catalyzed by the intein, but the final product rearrangement is a spontaneous, thermodynamically favored reaction that assures the irreversibility of protein splicing.

The first step in protein splicing is a reversible transition of the peptide bond between the amino end of the intein and its amino terminal flank (N-extein) into an ester or thioester bond. This transition depends on a nucleophilic attack of the bond by the side-chain of the Ser or Cys residues at the amino terminal end of the intein (-OH or -SH respectively). This reaction is termed N-O when the attacking atom is an Oxygen and N-S when this atom is Sulfur. This scheme shows the reaction with a Cys in the intein amino end. All inteins begin with either Ser or Cys residues,except for the two klbA inteins in M.jannaschii and Pyrococcus horikoshii OT3. These start with an Ala and if they are active it cannot be through this step since the Ala side-chain is a methyl group (CH3) not capable of nucleophilic attack.

In the second step the side-chain of the residue C-terminal to the intein (the first residue of the Carboxy (C) extein) attacks the ester (or thioester) bond at the amino end of the intein. Here too the attack is by a polar side chain of a Ser, Thr (both OH) or Cys (SH).This leads to a transesterification and formation of a branched intermediate with two amino ends, one of the N-extein and one of the intein. The intein is joined by peptide bond to the C-extein and the two exteins are joined by a thio/ester bond. This reaction is also reversible. All known inteins indeed have Ser, Thr or Cys directly following their C-end. This scheme shows the reaction with a Ser following the intein carboxy end.



Mechanism of Protein Splicing

In the third step the branched intermediate is resolved by the cyclization of the C-terminal intein residue. The intein is now fully excised from the N and C exteins that are yet linked to each other by the thio/ester bond. This step is ireversible driving the reaction forward. Almost all inteins have Asn as their carboxy end (as shown in this scheme) and its cyclization results in a succinimide ring. Two known inteins have Gln in their carboxy end and a variation of this step has been proposed to account for this. In brief, the reaction proceeds through Gln cyclization into a glutarimide ring.

The final steps consist of spontaneous shift of the thio/ester bond linking the exteins into a peptide bond (S/O-N acyl rearrangement) and probably some hydrolysis of the succinimide (or glutarimide) ring at the intein carboxy end to Asn and iso-Asn. These reactions are ireversible too and form the mature host protein, chemically identical to the product of an intein-less gene. Not much is biochemically known on the fate of the excised intein. In experiments where it is over-produced it seems to be rapidly degraded. However, genetic and phylogenetic analysis show that some inteins are also responsible for the homing of the intein gene into unoccupied intein integration points in homologous genes. This horizontal-transfer gene conversion is mediated by the homing endonuclease protein domain found in the central region of most inteins. The reaction is totally independent of the protein splicing reaction depicted here.

Inteins can be viewed as a class of highly unusual enzymes: (1) they catalyze three mechanistically distinct reactions; (2) they act on amino acid residues at their own

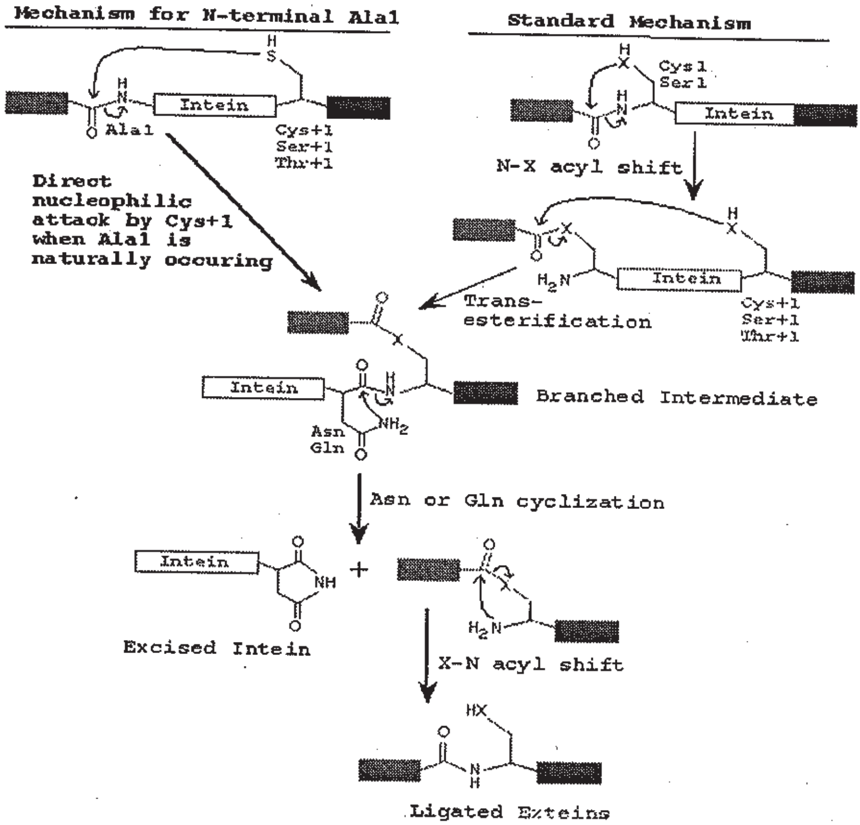
N- and C-termini, so that the intein enzymes are also their own substrate, analogous to the role of catalytic RNA in the self-splicing of group I introns; and (3) their catalytic center comprises the two extremities of a polypeptide chain, a situation which is rarely encountered in conventional enzymes and suggests an unusual protein structure. Our aim is to elucidate the mechanism of catalysis of protein splicing by defining the catalytic center of the self-splicing intein, both in terms of the amino acid side chains involved and their arrangement in 3-dimensional space.

Alternative Protein Splicing Mechanism

Variations in the intein-mediated protein splicing mechanism are becoming more apparent as polymorphisms in conserved catalytic residues are identified. Several families of inteins have been identified that begin with Ala rather than the consensus nucleophiles, Ser or Cys. In standard inteins, an N-terminal Ser, Cys or Thr is absolutely required for splicing. An N-terminal Ala cannot perform the initial reaction of the standard protein splicing pathway to yield the requisite N-terminal splice junction (thio)ester. However, experiments with the M. jannaschii KlbA intein demonstrated that Ala1 inteins can splice efficiently using an alternative protein splicing mechanism (Southworth 2000). In this non-canonical pathway, the C-extein nucleophile (Ser, Cys or Thr) attacks a peptide bond at the N-terminal splice junction rather than a (thio)ester bond, alleviating the need to form the initial (thio)ester at the N-terminal splice junction. The remainder of the two pathways is identical: branch resolution by Asn cyclization is followed by an acyl rearrangement to form a native peptide bond between the ligated exteins. Just like standard inteins, the Mja KlbA intein also requires the help of the conserved Thr and His in Block B to activate the N-terminal splice junction. We have also demonstrated splicing of the Mle DnaB intein (dnaB-b insertion site, E. Davis, M. Southworth & F. Perler, unpublished data) which is another Ala1 intein, suggesting that different families of naturally occurring Ala1 inteins should be capable of splicing.

The KlbA and Mle DnaB inteins have overcome the barriers to direct nucleophilic attack on the peptide bond at the N-terminal splice junction that are present in previously studied inteins with Ser or Cys at their N-terminus. It is unclear why other inteins can't perform similar reactions, since the Block B oxyanion hole is still available to facilitate direct attack on the N-terminal splice junction. Possibly, (thio)ester formation may be necessary in standard inteins to align the C-extein nucleophile, to remove steric hindrances or to induce a conformational shift that allows attack by the +1 nucleophile (Cys, Ser or Thr). The crystal structure of a S.cerevisiae VMA intein precursor has helped to resolve this question by revealing that Cys+1 is too far away to directly attack either a peptide or a thioester bond at the N-terminal splice junction, leading the authors to suggest that inteins must undergo a conformational shift to allow attack by the Cys+1 nucleophile (Poland 2000). We propose that Cys+1 (or its equivalent residue) in Ala1 inteins is already in position to attack the N-terminal splice junction amide bond in the precursors protein.

An Alternative Protein Splicing Mechanism for Ala1 Inteins

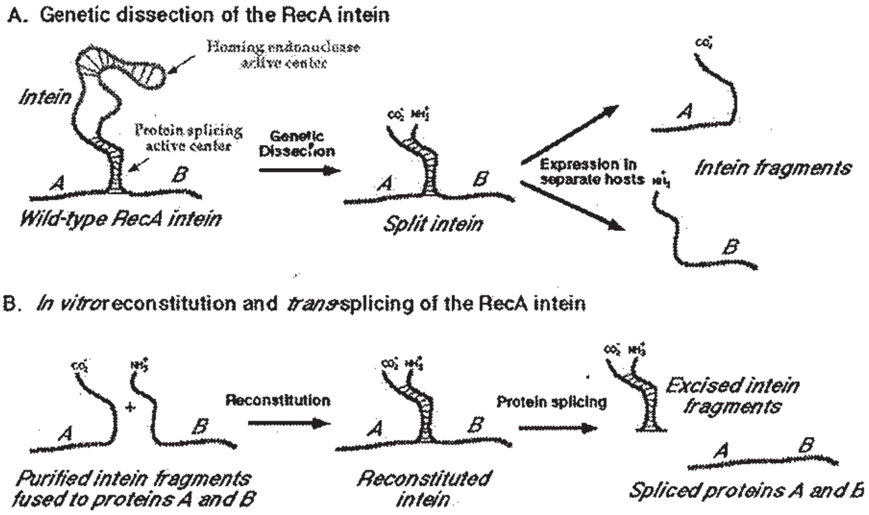


(X represents the sulfur or oxygen atom in Cys, Ser or Thr side chains)

A case study

The experimental system used in our studies is the intein from the RecA protein of Mycobacterium tuberculosis. As a first step in our investigation, we cloned the RecA intein between two affinity tags as artificial exteins and genetically dissected away the portions of the intein that are involved in its homing endonuclease function so as to generate a minimal protein splicing element. Further dissection of the protein splicing element into separate N- and C-terminal fragments (about 100 amino acids each) showed that protein splicing can also occur in trans. This allowed us to develop an efficient in vitro trans-splicing system in which purified N- and C-terminal intein fragments are reconstituted and allowed to undergo splicing. Under appropriate conditions, reconstitution and protein splicing can be studied separately, thus opening the way for analyzing both the structural and the catalytic basis of protein splicing.

Our experimental system for the study of protein splicing involves the RecA intein of M. tuberculosis, which plays a critical role in the repair of the DNA damage incurred by this pathogen when it invades the macrophages of its host. The elucidation of the mechanism by which the intein catalyzes protein splicing to yield active RecA protein from an inactive precursor may therefore suggest ways to control a process that plays an important role in the virulence of M. tuberculosis and may lead to the development of a new class of anti -mycobacterial drugs.


A. Genetic dissection of the RecA intein

B. *In vitro* reconstitution and *trans*-splicing of the RecA intein

Phylogeny

The phylogenetic distribution of inteins is sporadic. The presence of an intein in a particular gene does not necessarily mean that an extein homolog from a closely related species or strain will have the same intein. For example, look at the inteins present in the 3 insertion sites in DNA polymerases from various strains of archaea or GyrA inteins in various species of Mycobacterium. At this time, it is not clear whether this pattern of intein distribution represents loss of ancient inteins or more recent acquisition of inteins due to intein gene mobility. In many cases analyzed, the codon usage and GC content of the intein coding region is different from the surrounding extein coding region, suggesting recent horizontal transmission. Organisms that have a large number of inteins may have acquired this large number of inteins because they (1) can easily take up DNA from the environment (naturally competent), (2) share viruses, conjugative elements, plasmids, etc. that have broad host ranges, or (3) have very efficient gene conversion machinery, double-stand breaks repair systems and/or recombination systems.

Several inteins, such as the DnaB, RIR1, GyrA and Pol inteins, are present at the same extein insertion site in extein homologs from several species, including extein homologs in organisms from different phylogenetic domains. Perler 1997 suggested that inteins present in the same insertion site of an extein homolog be considered intein alleles or homologs. Inteins grouped by extein insertion site are tabulated in the Intein Alleles Section and extein insertion sites are named according to. Intein alleles are more closely related to each other than to other inteins in the same organism or even in the same gene.

Extein proteins may also have multiple inteins present at different insertion sites within the extein (for example, Tli Pol, Tsp-TY Pol, Mja RFC, Mja RNR or Pfu RIR1). A few proteins have 3 inteins.

Intein alleles are more closely related to each other than to other inteins because they are either descendants of an ancestral intein or have been recently mobilized into that site based on homing endonuclease specificity. Another way of saying this, is that inteins that have the same homing endonuclease specificity are more related to each other than to other inteins. Remember, the homing endonuclease recognition site determines the intein insertion site because the double-strand break made by the homing endonuclease initiates the site-specific gene conversion reaction leading to intein acquisition.

Applications

- Inteins are naturally occurring proteins that are involved in the precise cleavage and formation of peptide bonds in a process known as protein splicing.
- Genetic engineering has allowed the controllable cleavage of peptide bonds at either the N- or C-terminus of the intein.

- Inteins displaying controllable cleavage have been used in the isolation of bacterially expressed proteins possessing either a C-terminal thioester or an N-terminal cysteine.

- The specific placement of these reactive groups has allowed either protein-protein or protein-peptide condensation through a native peptide bond.

- This review describes the methods used to specifically generate these reactive groups on bacterially expressed proteins and some applications of this technique, known as intein-mediated protein ligation.

- Furthermore, a versatile two intein (TWIN) system will be described which enables the circularization and polymerization of bacterially expressed proteins or peptides.

- Construction of a mini-intein fusion system to allow both direct monitoring of soluble protein expression and rapid purification of target proteins.

- Affinity purification of recombinant proteins has been facilitated by fusion to a modified protein splicing element (intein).

- The fusion protein expression can be further improved by fusion to a mini-intein, i.e. an intein that lacks an endonuclease domain.

- In mammalian cells, protein-protein interactions constitute essential regulatory steps that modulate the activity of signaling pathways.

- In recent years, several approaches towards understanding the interactions have been developed.

- Protein splicing, the protein equivalent of RNA splicing, is a newly discovered posttranslational process that proceeds through a branched protein intermediate and produces two separate polypeptides from one gene.

References

1. Editorial: Nice splice. Nat Struct Biol 4(7):507-508. PubMed ID:9228936

2. Abel-Santos, E., Scott, C.P., Benkovic, S.J. (2003) Use of inteins for the in vivo production of stable cyclic peptide libraries in E. coli. Methods Mol. Biol. 205:281-294. PubMed ID:12491894

3. Adam, E. and Perler, F.B. (2002). Development of a Positive Genetic Selection System for Inhibition of Protein Splicing using Mycobacterial Inteins in Escherichia coli DNA Gyrase Subunit A. J. Molec. Microbiol. Biotech. 4:479-487.

4. Albert, A., Dhanaraj, B., Genschel, U., Khan, G., Ramjee, M. K., Pulido, R., Sibanda, B. L., von Delft, F., Witty, M., Blundell, T. T., Smith, A. G., and Abell, C. (1998). Crystal structure of aspartate decarboxylase at 2.2 A resolution: evidence for an ester in protein self-processing. Nat Struct Biol 5:289-293. PubMed ID:9546220

5. Beachy, P. A., Cooper, M. K., Young, K. E., von Kessler, D. P., Park, W., Hall, T. M. T., Leahy, D. J. and Porter, J. A. (1997) Multiple Roles of Cholesterol in hedgehog protein biogenesis and signaling. Cold Spring Harbor Symp. Quant. Biol. 62:191-204. PubMed ID:9598352

6. Becker, Y. (1998) Molecular evolution of viruses - Past and Present, Part 2 - An introduction. Virus Genes 16(1):7-11. PubMed ID:9562887

7. Camarero, J.A., Fushman, D., Sato, S., Giriat, I., Cowburn, D., Raleigh, D.P. and Muir, T.W. (2001). Rescuing a destabilized protein fold through backbone cyclization. J. Mol. Biol. 308(5):1045-1062. PubMed ID:11352590

8. Dalgaard, J. Z., Klar, A.J., Moser, M. J., Holley, W.R., Chatterjee, A. and Mian, I. S. (1997B) Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. Nuc. Acids Res. 25:4626-2638. PubMed ID:9358175

9. Daugelat, S. and Jacobs, W.R. Jr. (1999). The Mycobacterium tuberculosis recA intein can be used in an ORFTRAP to select for open reading frames. Protein Sci. 8:644-653. PubMed ID:10091667

10. Fitzsimons-Hall, M., Noren, C.J., Perler, F.B. and Schildkraut, I. (2002) Creation of an Artificial Bifunctional Intein by Grafting a Homing Endonuclease into a Mini-intein. J. Mol. Biol. 323:173-179. PubMed ID:12381313

11. Flavell, R.R., Huse, M., Goger, M., Trester-Zedlitz, M., Kuriyan, J. and Muir, T.W. (2002) Efficient Semisynthesis of a Tetraphosphorylated Analogue of the Type I TGF beta Receptor. Organic Letters 4(2):165-168.PubMed ID:11796041

12. Ghim, S.Y., Choi, S.K., Shin, B.S. and Park, S.H. (1998) An 8kb nucleotide sequence at the 3' flanking region of the sspC gene (184 degrees) on the Bacillus subtilis 168 chromosome containing an intein and an intron. DNA Res. 5:121-126. PubMed ID:9679200

13. Ghosh, I., Sun, L. and Xu, M.Q. (2001). Zinc Inhibition of Protein Trans-splicing and Identification of Regions Essential for Splicing and Association of a Split Intein. J. Biol. Chem. 276:24051-24058. PubMed ID:11331276

14. Hashimoto, H., Nishioka, M., Fujiwara, S., et al. (2001) Crystallographic structure of DNA polymerase from hyperthemophilic archaeon Pyrococcus kodakaraensis KOD1. J. Mol. Biol. 306:469-477. PubMed ID:11178906

15. He, Z., Crist, M., Yen, H., Duan, X., Quiocho, F. A., and Gimble, F. S. (1998). Amino Acid Residues in Both the Protein Splicing and Endonuclease Domains of the PI-Scel Intein Mediate DNA Binding. J Biol Chem 273:4607-15. PubMed ID:9468518

16. Ichiyanagi, K., Ishino, Y., Ariyoshi, M., Komori, K. and Morikawa K. (2000). Crystal Structure of an Archaeal Intein-encoded Homing Endonuclease PI-PfuI. J. Mol. Biol. 300:889-901. PubMed ID:10891276

17. Inoue, K., Demel, R., de Kruijff, B. and Keegstra, K. (2001). The N-terminal portion of the preToc75 transit peptide interacts with membrane lipids and inhibits binding and import of precursor proteins into isolated chloroplasts. Eur. J. Biochem. 268(14):4036-4043. PubMed ID:11453998

18. Kawashima, T., Yamamoto, Y., Aramaki, H., Nunoshiba, T., Kawamoto, T., Watanabe, K., Yamazaki, M., Kanehori, K., Amano, N., Ohya, Y., Makino, K. and Suzuki, M. (1999) Determination of the complete genomic DNA sequence of Thermoplasma volvanium GSS1. Proc. Jpn. Acad. 75:213-218.

19. Keeling, P. J. and Roger A. J. (1995) The selfish pursuit of sex [letter]. Nature 375:283. PubMed ID:7753189

20. Kumar, R. A., Vaze ,M. B., Chandra. N. R., Vijayan, M. and Muniyappa, K. (1996) Functional characterization of the precursor and spliced forms of recA protein of Mycobacterium tuberculosis. Biochemistry 35:1793-1802. PubMed ID:8639660

21. Li, Y., Zhao, Y., Hatfield, S., Wan, R., Zhu, Q., Li, X., McMills, M., Ma, Y., Li, J., Brown, K.L., He, C., Liu, F. and Chen, X. (2000). Dipeptide seryl-histidine and related oligopeptides cleave DNA, protein, and a carboxyl ester. Bioorg. Med. Chem. 8:2675-2680. PubMed ID:11131157

22. Li, H.H., Thomas, M.J., Pan, W., Alexander, E., Samuel, M., and Sorci-Thomas, M.G. (2001). Preparation and incorporation of probe-labeled apoA-I for fluorescence resonance energy transfer studies of rHDL. J. Lipid Res. 42(12):2084-2091. PubMed ID:11734582