ELSEVIER

# Recent advances in protein splicing: manipulating proteins *in vitro* and *in vivo*

Ming-Qun Xu and Thomas C Evans Jr

Protein splicing is an intricate self-catalyzed protein rearrangement that converts an inactive protein precursor to biologically active proteins. In the past decade, mechanistic studies and extensive engineering of the naturally occurring protein splicing elements, termed inteins, has led to the development of numerous novel technologies. These intein-based methodologies permit *in vitro* and *in vivo* protein processing in ways previously not possible using traditional biochemical and genetic approaches. Inteins have been utilized in the production of protein and peptide arrays, as molecular switches and in the reconstitution of functional proteins by split-gene techniques.

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

Corresponding author: Xu, Ming-Qun (xum@neb.com)

**Current Opinion in Biotechnology** 2005, **16**:440–446

This review comes from a themed issue on
Protein technologies and commercial enzymes
Edited by Bernhard Hauer and Brian K Kay

Available online 18th July 2005

0958-1669/$ – see front matter


**DOI 10.1016/j.copbio.2005.06.012**

## Introduction

Protein splicing is an extraordinary post-translational processing event that involves the precise removal of an internal polypeptide segment, termed an intein, from a precursor protein with the concomitant ligation of the flanking polypeptide sequences, termed exteins [1]. Since its discovery in 1990, more than 200 inteins have been identified in all three domains of life [2]. The inteins, ranging from 128 to 1650 amino acids, share a set of highly conserved sequence motifs. The majority of known inteins appear to be bifunctional, as they also contain the characteristic motifs of a homing endonuclease that confers genetic mobility upon the intein-encoding gene. An endonuclease insertion splits the region required for splicing. A small number of inteins lack an endonuclease-coding region and are termed mini-inteins. Of special interest are the naturally occurring *trans*-splicing inteins in which a host gene is split into two separate coding regions, each fused to either the N-terminal or C-terminal portion of an intein-coding region [3]. Formation of the full-length host protein occurs when the N-terminal and C-terminal intein regions come together to reconstitute protein splicing activity.

Here, we review some of the recent advances in protein splicing research and discuss a number of intein-based applications.

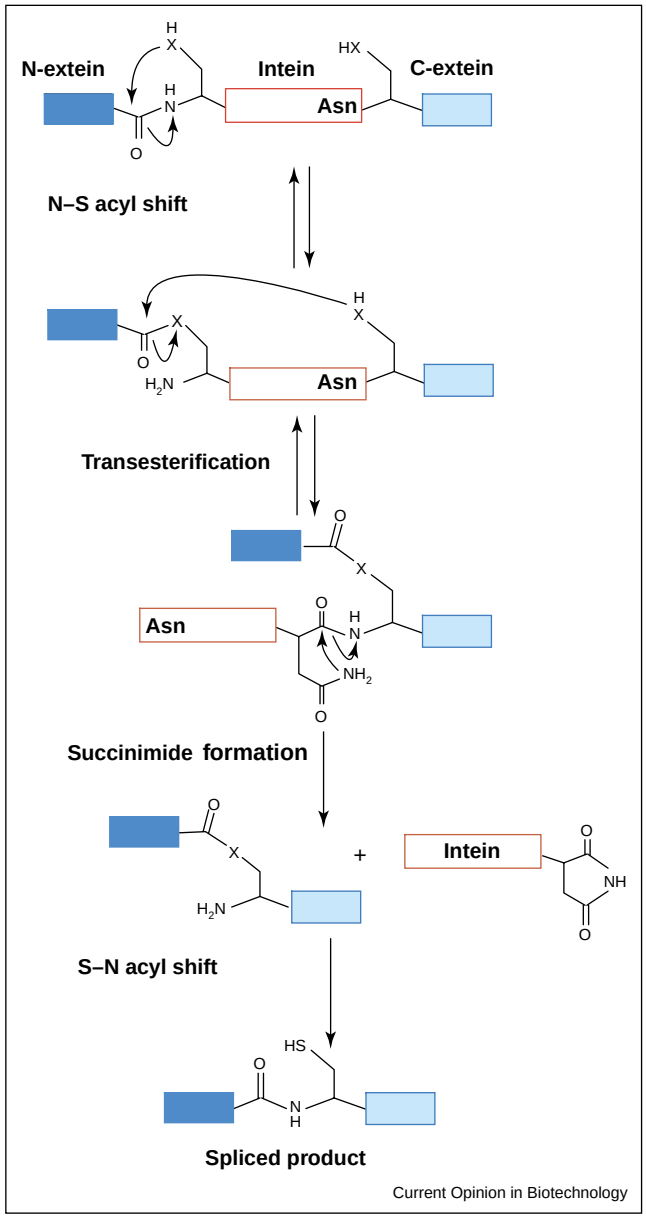
## Splicing mechanism and intein family

Many inteins have been shown to self-splice *in vitro* without the requirement of external energy or protein cofactors [4]. The mechanism of protein splicing has been elucidated by the identification of key catalytic amino acid residues and intermediates (Figure 1). Most inteins start with a cysteine or serine residue that is responsible for an acyl shift at the N-terminal splice junction. The first C-extein residue following the scissile bond at the C-terminal splice site is invariably a cysteine, serine or threonine and the sulfhydryl or hydroxyl group on their sidechain nucleophilicly attacks the linkage at the N-terminal splice junction, resulting in a branched intermediate. An asparagine residue typically precedes the C-terminal splice junction and is involved in the resolution of the branched intermediate by sidechain cyclization. Interestingly, a subfamily of inteins possessing an N-terminal alanine apparently initiates splicing by a direct attack on the peptide bond at the N-terminal splice junction by the sidechain of the first C-extein residue [5]. In addition, recently identified non-canonical inteins include those with glutamate or aspartate in place of the highly conserved C-terminal asparagine and bacterial intein-like proteins, mainly possessing a C-terminal glutamate, glycine or leucine [6–9]. A study of the oceanic nitrogen-fixing cyanobacterium *Trichodesmium erythraeum* revealed a remarkable intein organization showing the presence of three inteins (including one split intein) in the *dnaE* gene encoding the catalytic domain of DNA polymerase III [10]. The study of *T. erythraeum* has also led to the first report of the coexistence of multiple inteins and introns in a single gene [11,12]. A new example of a viral intein was recently found in Mimivirus [13]. The broad range of intein properties facilitates their use in diverse protein engineering strategies.

## Intein-mediated protein immobilization

The steps that underlie protein splicing consist of two acyl rearrangements, a transesterification and cyclization of an asparagine. The elucidation of the protein-splicing pathway led to the discovery that catalysis of each of the steps is often relatively independent. Formation of a thioester by an initial acyl rearrangement occurs even

**Figure 1**




Prototypical protein splicing mechanism. The hydroxyl- or sulfhydryl-containing sidechain of the intein N-terminal residue initiates an N–O or N–S acyl rearrangement, resulting in a (thio)ester linkage between the N-extein and C-extein. The sidechain of cysteine, serine or threonine at the C-terminal splice junction attacks the (thio)ester in a trans(thio)esterification reaction to form a branched intermediate. The intein is excised from the branch by cyclization of the intein C-terminal asparagine coupled to peptide bond breakage. A spontaneous O–N or S–N shift generates a native peptide bond. The X represents either an oxygen or sulfur atom.

when the subsequent steps are blocked by amino acid replacements (e.g. by substitution of the intein C-terminal asparagine with alanine) at the downstream splice junction [14,15]. Thioester formation is the basis for an intein-mediated purification system in which a target protein is fused to the N terminus of an intein and can be released in a thiol-induced reaction. This intein fusion system has been extended to produce recombinant proteins possessing a C-terminal thioester for ligation with synthetic peptides or recombinant proteins carrying a variety of modifications or chemical moieties [16,17]. Researchers have used this technique to incorporate various probes in a site-specific manner or to produce proteins with isotopically labeled regions for functional and structural analysis [18,19••]. Semisynthetic DNA–protein conjugates were also generated by use of a C-terminal thioester on an expressed protein and cysteinyl oligonucleotides [20,21•].

Recently, intein-mediated protein ligation has been further employed to generate protein or peptide arrays by improving binding efficiency and orientation of the target molecules. A technique was developed for the site-specific attachment of C-terminal biotinylated proteins onto avidin-coated glass slides (Figure 2a) [22•,23]. Similarly, single-chain antibodies expressed in *Escherichia coli* can be labeled for chip-based screening [24]. The immobilization of site-specifically oriented proteins might help to retain their biological activities. In addition, an extremely strong avidin–biotin linkage is beneficial to withstand various assay conditions. This scheme permits the arrays to be utilized for quantitative analysis because each target protein carries only one reactive site for the biotinylated tag, which in turn is capable of binding to avidin ligand immobilized on a glass slide.

Furthermore, a new strategy was recently demonstrated for making peptide arrays on low-cost nitrocellulose. This approach employs the intein-mediated protein ligation of synthetic peptide substrates to an intein-generated carrier protein (Figure 2b) [25•]. This method is intended to provide a simple solution to the problems associated with the variable binding of small peptide substrates to matrices. The commonly used method of synthetic peptide arrays on membrane support (SPOT synthesis) produces an excessive amount of peptide and therefore has limitations in peptide quantification and normalization. As intein-generated carrier proteins play a dominant role in binding and each carrier protein molecule has precisely one reactive site for a peptide possessing an N-terminal cysteine, the amount of peptide arrayed onto a membrane can be effectively normalized. This technique, termed intein-mediated peptide array has been applied to antibody characterization, epitope scanning and kinase assays, and resulted in an increase in sensitivity up to $10^4$-fold.

## Putting *trans*-splicing to work

In protein *trans*-splicing, a target gene is split into two segments and each half is fused to either the N-terminal or C-terminal portion of an intein-coding sequence; the two halves of the intein-coding sequence are not linked in