2014

Split Inteins: From Mechanistic Studies to Novel Protein Engineering Technologies

Neel H. Shah

SPLIT INTEINS: FROM MECHANISTIC STUDIES TO NOVEL PROTEIN ENGINEERING TECHNOLOGIES

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Neel H. Shah

June 2014

SPLIT INTEINS: FROM MECHANISTIC STUDIES TO NOVEL PROTEIN ENGINEERING TECHNOLOGIES

Neel H. Shah, Ph.D.

The Rockefeller University 2014

Inteins are auto-processing protein domains that carry out a post-translational process known as protein splicing. This process is characterized by excision of the intein (*int*ervening prot*ein*) domain from within a larger polypeptide sequence with concomitant ligation of the flanking extein (*ext*ernal prot*ein*) regions through a native peptide bond. Remarkably, a small subset of all inteins are naturally transcribed and translated as two fragments that efficiently associate and carry out the same biochemical process *in trans*, and these split inteins are potentially powerful tools for protein engineering. Recently, a split intein from the cyanobacterium *Nostoc punctiforme* (Npu) was discovered that can carry out protein splicing with a half-life of one minute, as opposed to hours as seen for previously characterized split and contiguous inteins. Inspired by the apparent uniqueness of this "ultrafast" splicing activity and its practical implications, we characterized several orthologous split inteins from the same family as Npu. Surprisingly, many of these inteins splice as quickly as Npu, and biochemical characterization of this family divulged sequence-activity correlations that provided insights into the molecular determinants for fast protein *trans*-splicing. Importantly, several of these inteins are extraordinarily efficient in their first auto-processing step, peptide bond cleavage coupled to thioester formation. Harnessing this property, along with efficient fragment association, a

streamlined iteration of Expressed Protein Ligation (EPL), the most prevalent protein semi-synthesis technique, was developed. Further insights into protein splicing were obtained by the development of a novel kinetic assay that allowed for quantitative observation of a crucial intermediate in the protein splicing pathway, the branched intermediate (BI). Using this assay, BI resolution was unambiguously identified as the rate limiting step for Npu splicing. Furthermore, the roles of extein residues in individual steps along the splicing pathway were teased apart. Using protein semi-synthesis, kinetic measurements, and structural techniques, C-extein composition was found to be intimately linked to active-site dynamics and BI resolution kinetics. In addition to chemical reactivity, the fragment assembly of Npu was also characterized. Mutation of charged residues at the binding interface demonstrated that split intein binding affinity was dominated by intermolecular electrostatic interactions. By swapping charged residues between the intein fragments, a new split intein was engineered with orthogonal binding and reactivity to the wild-type Npu split intein. The wild-type and charge-swapped inteins could be used in protein semi-synthesis endeavors requiring parallel selective splicing reactions in one pot. Finally, using a combination of biophysical techniques, the mechanism of split intein assembly was elucidated. Our analyses indicated that the assembly follows a unique trajectory comprised of coupled binding and folding of disordered regions of each fragment followed by a collapse of the structure into a stable functional domain. Collectively, these structural and functional studies not only provide insights into the inner workings of inteins but will also continue to aid in the development of important protein engineering technologies.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Professor Tom Muir, for giving me the opportunity to work in his lab, for mentoring me through several projects with the right balance of guidance and freedom, and for teaching me how to do rigorous science.

I would like to thank the members of my Faculty Advisory Committee, Professors Seth Darst and Rod MacKinnon, for their helpful suggestions throughout my thesis work and for their honest assessment of my research progress and future research plans. I would also like to thank Professor Dan Raleigh from Stony Brook University for serving as the external examiner of my thesis defense.

Throughout graduate school, I have had the opportunity to work with and learn from several outstanding researchers that I wish to thank. I was initially trained in the Muir lab by Dr. Steve Lockless, who taught me basic molecular biology techniques and gave me a unique perspective on biological sciences that has stuck with me ever since. After working with Steve, I effectively learned every experimental biochemistry technique that I know from Dr. Miquel Vila-Perelló. Miquel taught me how to make peptides and proteins, how to be a rigorous analytical chemist, and how to design good experiments. Furthermore, his patience allowed him to seamlessly traverse the boundaries between mentor, collaborator, and friend. I would like to thank Miquel for working with me on several projects, being my sound-board even when we weren't doing experiments together, and for getting coffee with me almost every day for the past several years.

I am grateful to the entire Muir lab for their support and loyalty. As a group, we have all had the good fortune to work in a friendly environment where nobody hesitates to help one another and everybody brings something unique to the table, both in scientific knowledge and in personality quirks. In particular, I want to thank Dr. Lenka Bittova, our fearless lab manager ever since our move to Princeton University, for keeping the lab running and fixing every little problem we encountered. I also wish to thank all of the post-doctoral fellows in the lab for sharing their expertise with me when I asked, all the while treating me as a peer. I want to recognize Adam Stevens, who is enthusiastically carrying on my unfinished intein business while pursuing his own curiosities.

Over the past two and a half years, I have had the opportunity to collaborate with Professor David Cowburn's lab at the Albert Einstein College of Medicine, specifically with Dr. Ertan Eryilmaz. I want to thank David for his willingness to pursue any project rooted in basic biophysical curiosities. I especially want to acknowledge Ertan. He is one of the most diligent scientists I have ever worked with, and our collaboration has been constantly exciting, educational, and fruitful.

I wish to acknowledge the Dean's office at The Rockefeller University and Megan Krause in the Department of Chemistry at Princeton University for their administrative assistance. In particular, their help during the transition from Rockefeller to Princeton made this process as smooth and painless as possible.

Finally, I would like to thank my family for their unconditional love and support. I especially want to thank my wife, Jen, for listening to my incessant complaining and for getting excited with me when I was ecstatic about even the most menial successes.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 4

Chapter 5

Chapter 6

Chapter 7

LIST OF TABLES

Chapter 1: Introduction

The diversity of proteins found in nature is immediately apparent from the vast array of biochemical functions carried out by these proteins to sustain living organisms. To a first approximation, these functions are dictated by a protein's primary amino acid sequence, which is transcribed and translated from the gene encoding that protein. This sequence carries all of the necessary information for a newly synthesized protein to fold into a well-defined three-dimensional structure, and this structure in turn confers its function.[1] In reality, however, many proteins require additional factors to fold into their active conformation,[2] while others remain intrinsically disordered as a requirement for their function.[3] Furthermore, most proteins are matured, activated, inhibited, translocated, and/or degraded through the chemical modification of their side chains and backbones after protein synthesis, adding yet another layer of complexity to their structure and function.[4]

In many cases, these post-translational modifications (PTMs) expand the chemical and structural repertoire of the canonical twenty amino acids by the addition of new functional groups to their side chains. These enzymatically applied modifications include (but are not limited to) phosphorylation, acetylation, methylation, lipidation, glycosylation, and hydroxylation (Figure 1.1a), and the dynamic interplay between their addition and removal governs biological signaling. In other cases, the primary sequence of a protein is post-translationally altered by the scission of one or more peptide bonds. This "processing" of a polypeptide chain is often carried out enzymatically by proteases (Figure 1.1b) and is most commonly utilized for protein degradation. However, it can also serve to activate an enzyme (e.g. the cleavage of prothrombin to yield thrombin),

remove a translocation signal (e.g. signal peptide removal during antibody secretion), or mature a protein into a functional state (e.g. the conversion of proinsulin into active insulin). Remarkably, several classes of proteins can modify their own peptide backbones, and these modules are referred to as auto-processing domains (Figure 1.1c).

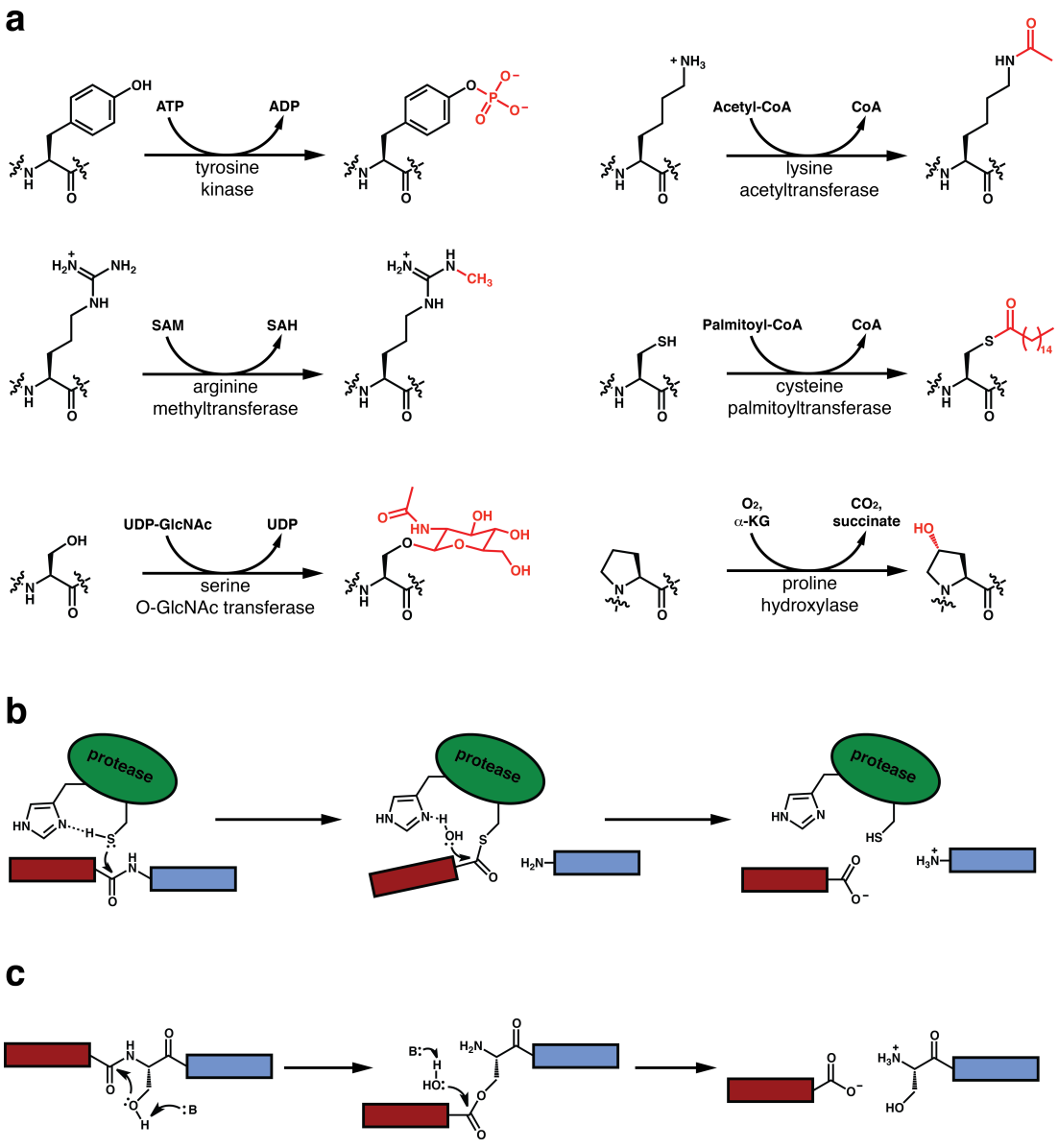


Figure 1.1. Post-translational modification of proteins. a. Examples of functional group additions enzymatically applied to various proteinaceous amino acids. **b.** Proteolytic cleavage of a peptide bond by a cysteine protease. **c.** Auto-processing of a peptide bond to yield two hydrolyzed polypeptide fragments (B refers to a base, typically an amino acid within the auto-processing domain or an activated water molecule).

The capacity for proteases and auto-processing domains to modify polypeptide sequences has garnered tremendous interest, not only for its biological significance, but also for its practical utility. Indeed, biochemists regularly use proteases to remove affinity purification tags from recombinant proteins[5] and to process complex biological mixtures for proteomics experiments.[6] Proteases are also commonplace in industrial settings, where they are used in detergents and for food production.[7] While auto-processing domains are less prevalent in technological settings, some of these proteins are rapidly emerging as powerful tools for chemical biology.[8] These useful auto-processing domains comprise a conserved family of proteins known as inteins.

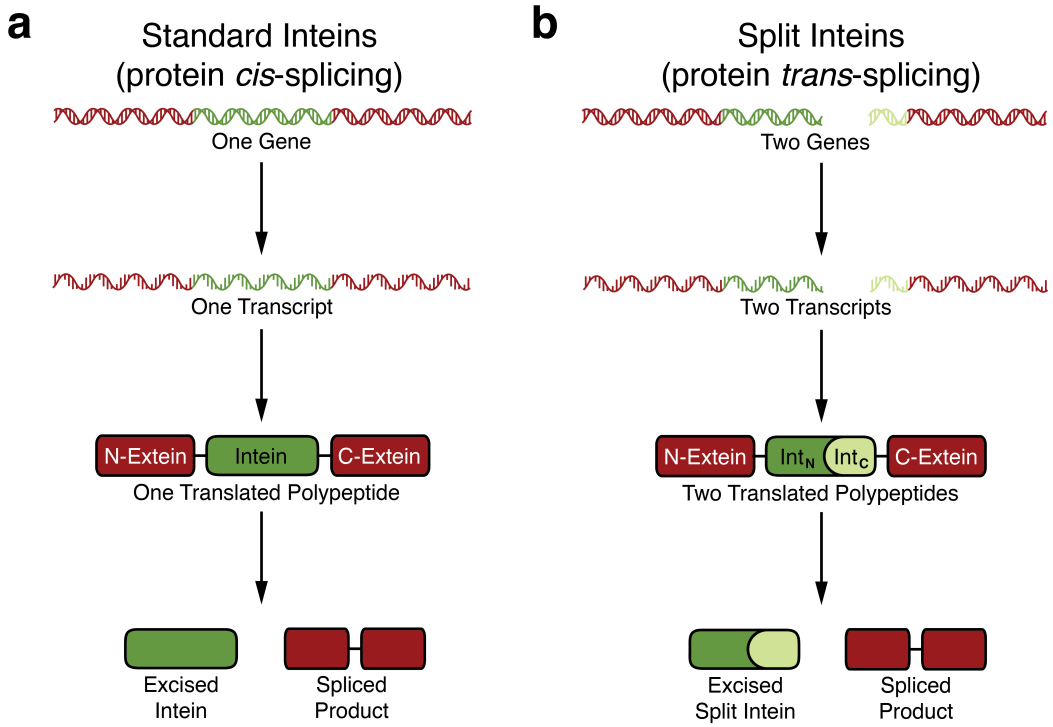


Figure 1.2. Protein splicing *in cis* and *in trans*. a. Protein *cis*-splicing by the more prevalent contiguous inteins. **b.** Protein *trans*-splicing by the rarer split inteins. Int_N refers to the N-intein and Int_C refers to the C-intein.

1.1. Protein splicing: a wide-spread post-translational modification

An intein (*in*tervening pro*tein*) carries out a unique auto-processing event known

as protein splicing in which it excises itself out from a larger precursor polypeptide through the cleavage of two peptide bonds and, in the process, ligates the flanking extein (*ex*ternal pro*tein*) sequences through the formation of a new peptide bond. This rearrangement occurs post-translationally (or possibly co-translationally), as intein genes are found embedded in frame within other protein-coding genes (Figure 1.2a). Furthermore, intein-mediated protein splicing is spontaneous; it requires no external factor or energy source, only the folding of the intein domain.

1.1.1. The vast phylogeny of inteins

In 1990, the first protein splicing domain was found embedded within the vacuolar proton-translocating ATPase gene in *Sacchromyces cerevisiae*.[9] Since then, taking advantage of several highly conserved sequence motifs in inteins, bioinformatic approaches have identified at least 600 putative intein genes in the genomes of unicellular organisms from all three domains of life (archaea, bacteria, and eukaryota) as well as several viral genomes (Figure 1.3).[10] Interestingly, this broad distribution of inteins pertains not only to the array of organisms in which they are found, but also the large variety of host genes in which they are embedded. To date, there are at least 70 different intein alleles, distinguished not only by the type of host gene in which the inteins are embedded, but also the integration point within that host gene.[10,11] Furthermore, some proteins have been found containing as many as three inteins embedded at different integration points within their gene,[12] and several organisms have more than one intein (as many as 19 inteins) in their genome (Figure 1.3).[10]

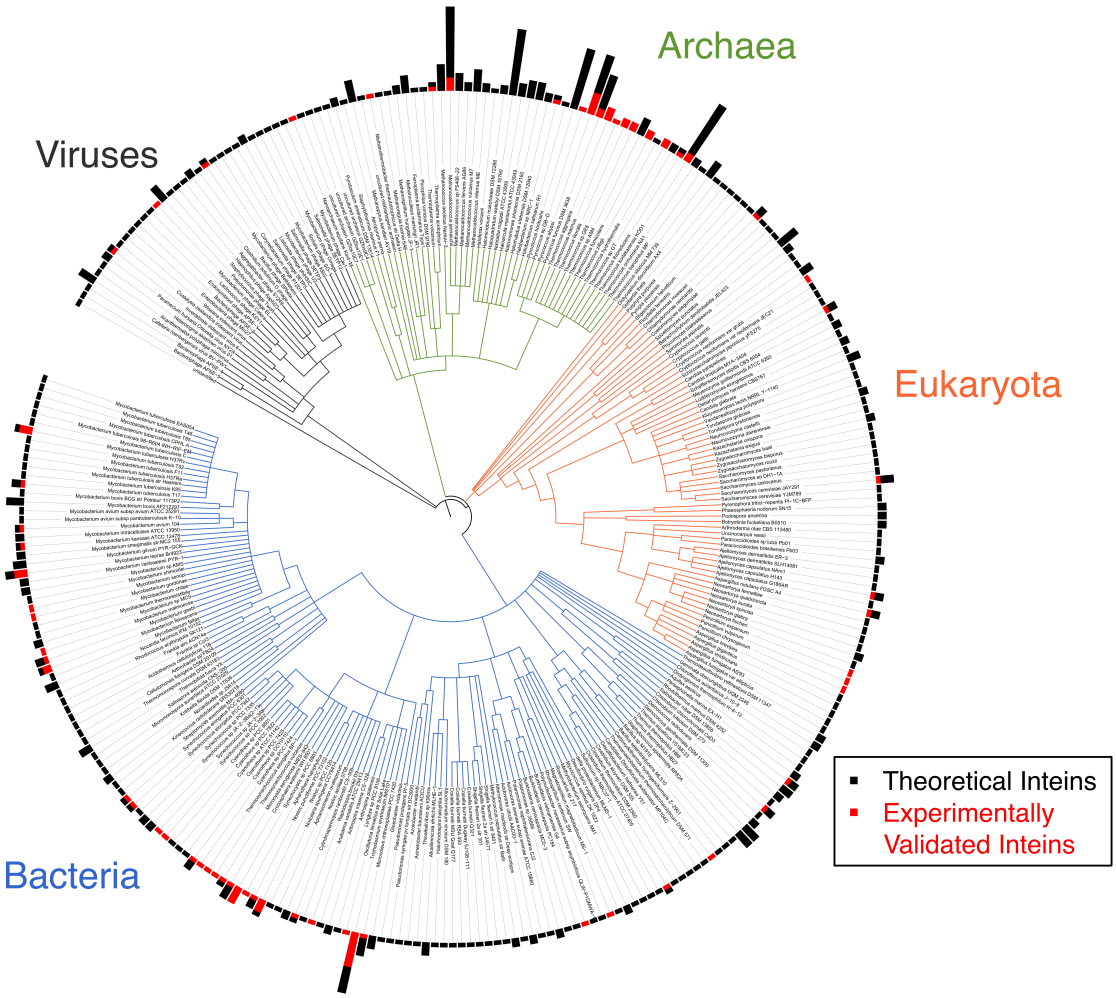


Figure 1.3. Phylogenetic distribution of intein-containing organisms. Roughly 300 organisms containing one or more intein are shown in this phylogenetic tree. The bars at the periphery of the tree denote the number of inteins in each organism. The smallest bar indicates one intein, and the largest bar indicates 19 inteins. Black bars indicate inteins identified based on their gene sequence whose splicing capacity has not yet been determined experimentally. Red bars indicate inteins that have been shown experimentally to facilitate protein splicing. Data was extrated from the NEB InBase[10] and the phylogenetic tree was generated using the Interactive Tree of Life online tool.[13]

Inteins are typically embedded within essential proteins involved in DNA replication, transcription, and maintenance (e.g. DNA or RNA polymerase subunits, DNA helicases, DNA gyrases, and ribonucleotide reductase) or in other housekeeping genes including essential proteases and metabolic enzymes.[11] Within these proteins, the auto-processing domains are commonly inserted at conserved sites in their host that are

crucial for host protein function (e.g. ligand binding sites or enzyme active sites).[14] Given this, it is largely believed that intein excision is required for host protein function, however no intein has been shown to have a clear regulatory role on its host protein or provide a fitness benefit to the host organism.[15,16] These facts beg the question, "Why do inteins exist?" Thus far, the prevailing view is that inteins are selfish genes with no obvious biological role. (This will be discussed further in Chapter 6). Regardless, their mere presence and persistence in microbial genomes is intrinsically fascinating.

1.1.2. Biological mechanisms of intein spread, persistence, and loss

The broad phylogenetic distribution of inteins suggests that these molecules have ancient origins. Nevertheless, for the reasons outlined below,[11] it is clear that their prevalence must not only be due to vertical, but also horizontal gene transmission: (1) Inteins are integrated into a wide variety of host proteins. (2) When an intein exists in one microbial host gene, it is not always found in an orthologous gene in a closely related organism. (3) Inteins are completely absent from multicellular organisms (although some multicellular organisms have homologous domains with related biochemical functions, discussed in section 1.3). (4) Allelic intein genes typically have higher homology and different codon usage than their host genes. These facts suggest that mechanisms exist to propagate inteins from one host gene and host organism to another, and that inteins can also be lost.

Many inteins have, inserted within their auto-processing domain, another functional module called a homing endonuclease domain (HED). HEDs make double-stranded DNA breaks at specific recognition sequences encoded within intein/HED-free

alleles of their own host genes. These breakages initiate a recombination process that results in the integration of an intein/HED gene into a previously intein/HED-free version of that gene (Figure 1.4a).

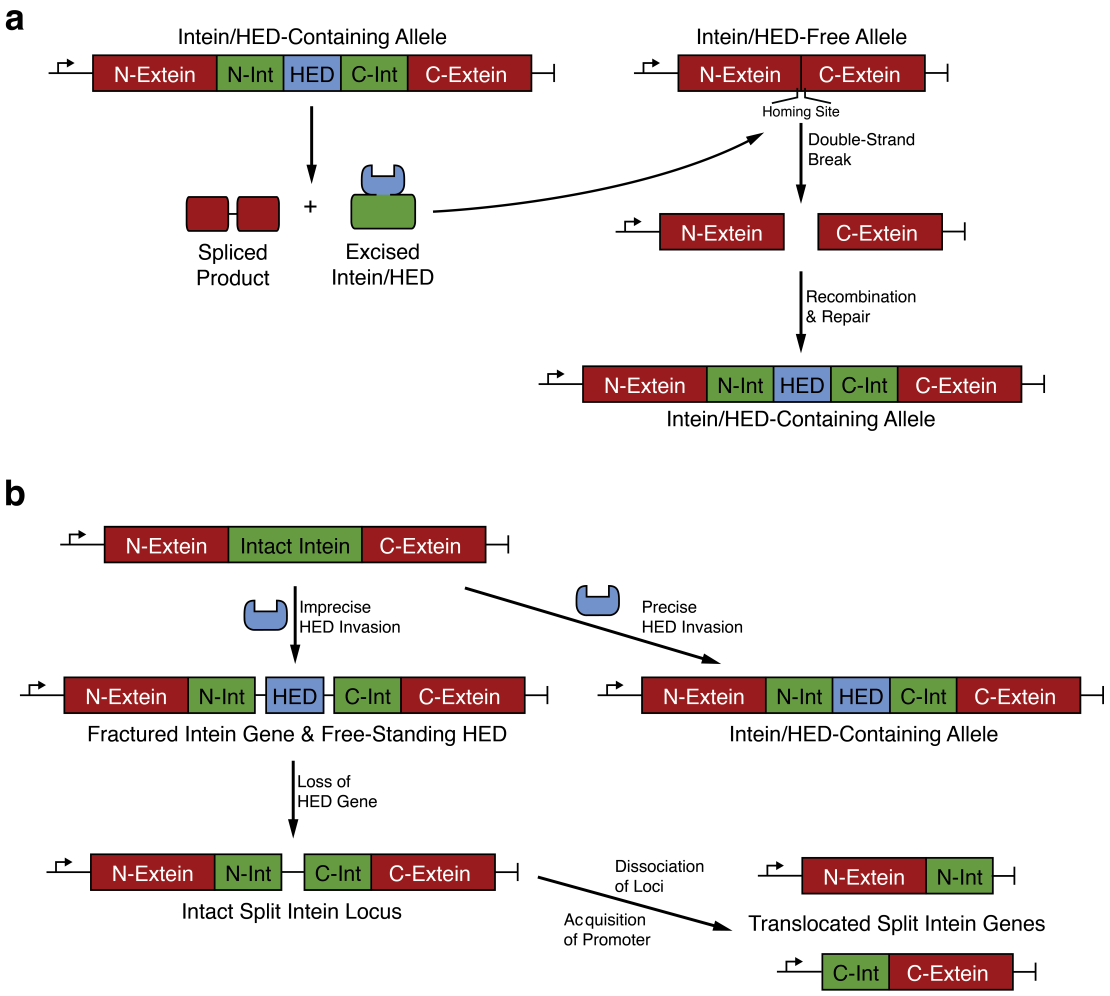


Figure 1.4. Intein spreading and splitting through homing endonucleases. a. Homing endonuclease activity to convert an intein-free allele into an intein-containing allele. **b.** Proposed mechanism for intein splitting due to aberrant homing endonuclease invasion followed by chromosomal rearrangements.[17]

This mechanism likely explains both intra- and inter-species intein spread involving the same host gene and insertion site.[11] However, given the sequence specificity of HEDs, it is unlikely that this mechanism allows for intein spread to different insertion sites within the same gene or different genes. Currently, it is unclear

how this process occurs, however it is noteworthy that non-allelic inteins have extremely low sequence homology when compared to allelic inteins.[11] This suggests that intein spread to a variety of host genes was an ancient process and that non-allelic inteins have diverged since this initial event. The recent discovery of inteins in viral genomes may hold the explanation to the early proliferation of inteins.[18]

The loss of an intein from a host gene may be driven by negative selection, if the intein is detrimental to host fitness. However, intein loss through gene deletion should be challenging, as it requires precise removal of the intein from within an essential gene. Presumably this process is more likely in diploid or polyploid organisms, where an intein-containing copy of a gene is expendable in the presence of an intein-free copy. Furthermore, the capacity for interchromosomal recombination provides another route to intein loss and may explain the dearth of inteins in multicellular organisms.

1.1.3. The emergence of split inteins

A small fraction (less than 5%) of the identified intein genes encode split inteins.[10] Unlike the more common contiguous inteins, these are transcribed and translated as two separate polypeptides, the N-intein and C-intein, each fused to one extein. Upon transcription and translation, the intein fragments spontaneously and non-covalently assemble into the canonical intein structure to carry out protein splicing *in trans* (Figure 1.2b).

Although several lineages of split inteins independently emerged during evolution, as evidenced by their divergent sequences and their insertion in at least five different host proteins,[17,19,20] the precise mechanism of intein splitting is not clear.

Interestingly, the split site of most split inteins is also the homing endonuclease insertion site within many contiguous inteins. In fact, some split intein genes are separated by an out-of-frame, free-standing HED gene, suggesting that aberrant insertion of an HED into an intein gene could fracture that gene (Figure 1.4b).[17] Oddly, the largest known family of split inteins, found within the DnaE genes of at least 20 cyanobacterial species, has unconserved genomic architecture.[19] In different cyanobacterial species, the intein fragments are located in dramatically different regions of the bacterial chromosome, and in some cases, the fragments are encoded on opposite strands as well. This suggests that after an initial fracturing of the intein gene in an early cyanobacterium, the resulting locus was unstable and further rearranged as the organism speciated (Figure 1.4b).

Split inteins present unique evolutionary challenges and opportunities for their host organisms. On one hand, splitting of an intein gene should be effectively irreversible, and the resulting gene fragments, if still transcribed and translated, have to contend with splicing their exteins *in trans*. These constraints provide a strong selection filter that would either lead to the termination of that lineage or significant optimization of the newly split intein. (The latter will be discussed in Chapter 5.) On the other hand, these split genes could provide a way to regulate the host gene's activity. Furthermore, it has been postulated that if an organism contains multiple cross-reactive split inteins, they could serve as a platform for protein evolution through domain shuffling.[21] (These points will be discussed further in Chapter 6.)

1.2. The chemical mechanism of protein splicing

Given that inteins are wide-spread in nature, it is not surprising that non-allelic

inteins have low sequence homology (<40%). Despite this fact, inteins are unified by their common biochemical mechanism for protein splicing, which is carried out by several conserved sequence motifs distributed throughout their primary amino acid sequences (Figure 1.5a).

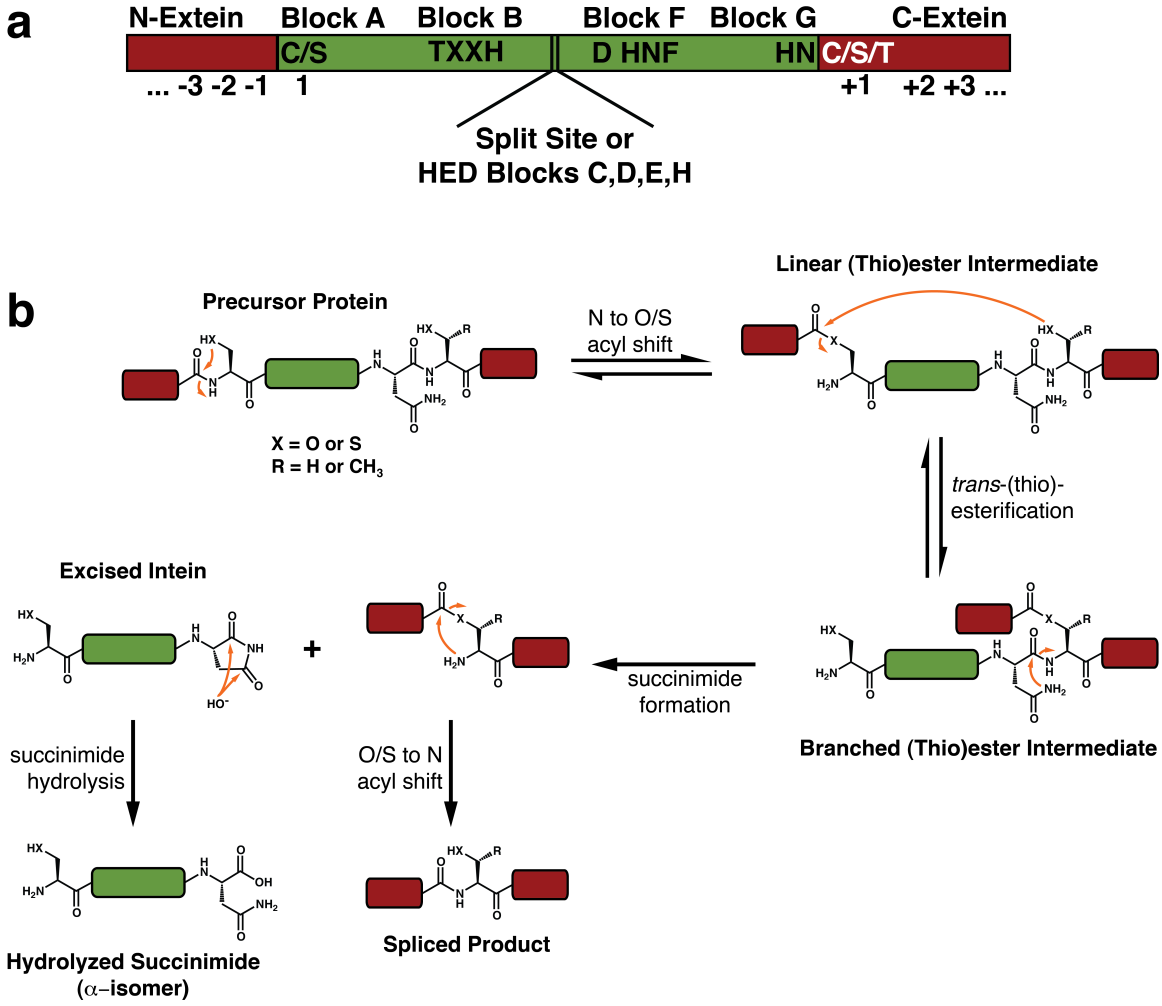


Figure 1.5. The chemical mechanism of protein splicing. a. Conserved sequence motifs that facilitate protein splicing. **b.** The canonical mechanism of protein splicing.

1.2.1. The canonical mechanism and the conserved sequence motifs

The mechanism of protein splicing entails a series of a acyl-transfer reactions that result in the cleavage of two peptide bonds at the intein-extein junctions and the formation of a new peptide bond between the N- and C-exteins (Figure 1.5b).[22] This

process is initiated by activation of the peptide bond joining the N-extein and the N-terminus of the intein. Virtually all inteins have a cysteine or serine at their N-terminus (Block A) that attacks the carbonyl carbon of the C-terminal N-extein residue. This N to O/S acyl-shift is facilitated by a conserved threonine and histidine found in Block B (also referred to as the TXXH motif), along with a commonly found aspartate in Block F, and results in the formation of a linear (thio)ester intermediate. Next, this intermediate is subject to *trans*-(thio)esterification by nucleophilic attack of the first C-extein residue (+1), which is invariably a cysteine, serine, or threonine. The resulting branched (thio)ester intermediate is resolved through a unique transformation: cyclization of the highly conserved C-terminal asparagine of the intein. This process is facilitated by the Block F histidine (found in a highly conserved HNF motif) and the penultimate Block G histidine and may also involve the Block F aspartate. This succinimide formation reaction excises the intein from the reactive complex and leaves behind the exteins attached through a non-peptidic linkage. This structure rapidly rearranges into a stable peptide bond in an intein-independent fashion. In addition, the excised intein succinimide will slowly hydrolyze into an α- or β-isomer of the carboxylic acid.

1.2.2. Variations on the canonical mechanism

Some inteins have slightly divergent sequences in the canonical splicing motifs which require alternate mechanisms of splicing. Perhaps the most dramatic of these differences is the lack of a nucleophilic cysteine or serine in Block A. Inteins lacking a Block A nucleophile commonly have an alanine or proline as their N-terminal residue and cannot initiate splicing through the formation of a linear (thio)ester intermediate.

Remarkably, these inteins either directly form the typical branched intermediate upon N-terminal activation[23] or proceed through a different branched thioester intermediate using an unique cysteine within Block F before forming the canonical branched structure.[24] The capacity to bypass the linear (thio)ester intermediate in some inteins is not surprising, as several studies have shown that the N-terminal scissile peptide bond in inteins is destabilized or twisted in the precursor protein.[25,26] Furthermore, the structure of the *Mja*KlbA intein, which contains an N-terminal alanine, shows that this peptide bond is in a *cis* conformation,[27] which may be unstable and susceptible to nucleophilic attack.

In another surprising variation on the splicing mechanism, a few inteins have been found ending in a glutamine or aspartate, rathern than asparagine.[28] These inteins could theoretically proceed through the same chemical mechanism, however glutamine cyclization should be less favorable than asparagine cyclization, as it would proceed through a six-membered, rather than five-membered, cyclic intermediate. Cyclization of aspartate would be geometrically similar to asparagine and proceed through the formation of a succinic anhydride rather than a succinimide.

Perhaps the most subtle but intriguing deviation from the canonical splicing motifs is the lack of a penultimate histidine.[29] In the cyanobacterial split DnaE inteins, this Block G histidine is either a serine or alanine.[19] In most inteins, both the Block F and Block G histidines are crucial for resolution of the branched intermediate, as they work in concert to activate the asparagine nucleophile then to protonate the forming amine.[30] Thus it is surprising that divergent inteins can complete the splicing reaction. Currently, it is not clear how these inteins circumvent the need for two histidines, however other proximal amino acids may serve as surrogate general acids or bases during splicing.

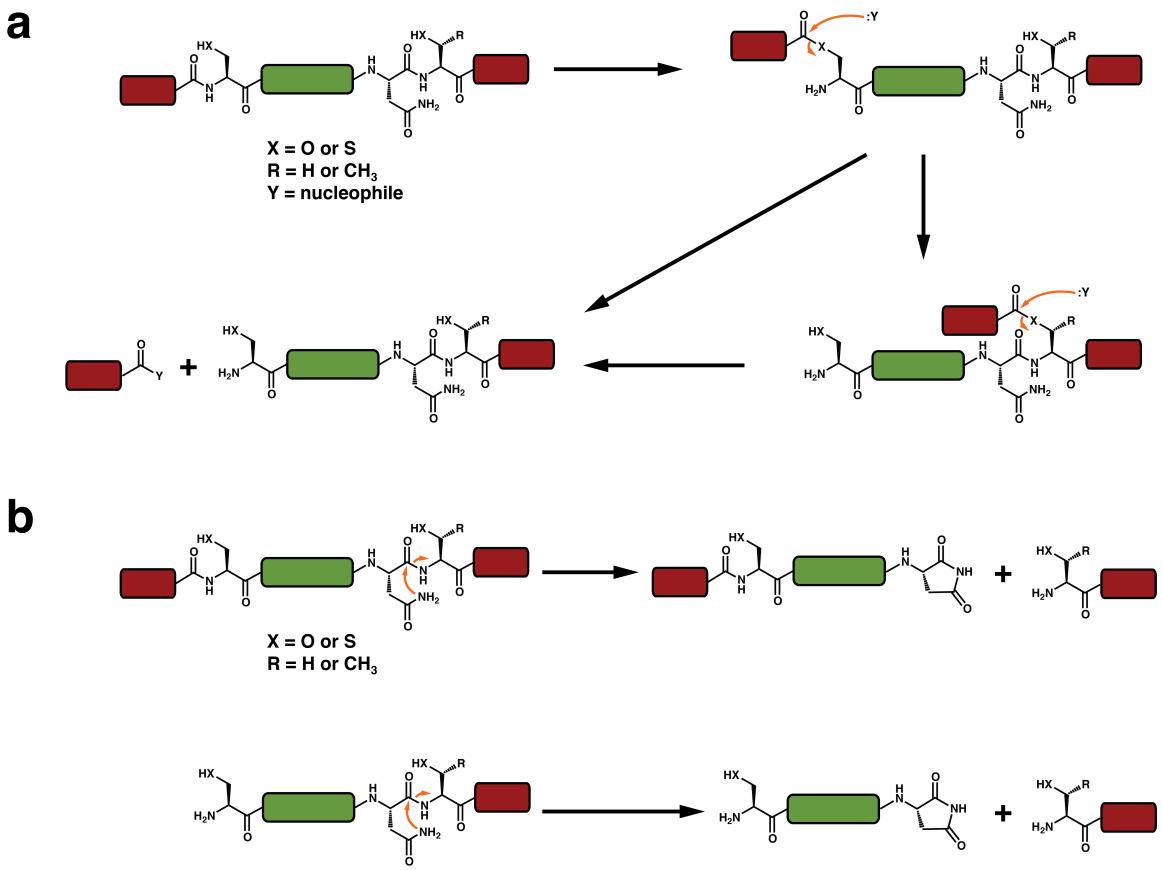


Figure 1.6. Side reactions during protein splicing. a. N-extein cleavage, also known as N-terminal cleavage, from the linear or branched (thio)ester intermediate. **b.** C-extein cleavage, also known as C-terminal cleavage, from the precursor protein or the N-extein cleaved protein.

1.2.3. Side reactions during protein splicing

Two common side reactions have been observed during experimental analysis of protein splicing reactions.[31] The first, N-extein cleavage (also known as N-terminal cleavage) can occur when an external nucleophile, commonly water or a thiol, attacks the linear or branched (thio)ester intermediates to yield a free N-extein (Figure 1.6a). The second, C-extein cleavage (also known as C-terminal cleavage) can occur from the precursor protein or the N-extein cleaved protein, when the C-terminal asparagine cyclizes in the absence of the branched intermediate structure, releasing the C-extein.

While both of these reactions can be enhanced by mutation of critical intein residues and have been exploited for various technological purposes,[31-34] it is not clear whether these reactions are prevalent in an intein's native environment. Interestingly, however, intein-related auto-processing domains have been found in nature that exploit these side reactions for biochemical outcomes other than protein splicing.

1.3. Hint domains: one fold, many functions

Inteins are actually part of a larger class of proteins known as Hedgehog/intein (Hint) domains. This superfamily is comprised of three members: inteins, bacterial intein-like (BIL) domains, and Hedgehog auto-processing (Hog) domains. While different Hint domains have various insertions, including the large homing endonuclease domains found in some inteins, all of these proteins have a the same core fold comprised primarily of several two- or three-strand β-sheets and loops along with two short α-helices (Figure 1.7a-g). Interestingly, this conserved horseshoe-like core has a pseudo two-fold symmetry. Given this symmetry, it is believed that the Hint fold likely arose from the gene duplication event of some progenitor protein with unrelated function.[35]

The most telling feature of the Hint fold is that its complex topology results in the presentation of the N- and C-termini in close proximity. Not surprisingly, all of the conserved sequence motifs found in inteins that confer their activity (Figure 1.5a) surround these termini (Figure 1.7h). The BIL domains and Hog domains carry out biochemical processes related to protein splicing, and thus they have retained one or more of these canonical sequence motifs near the active site.

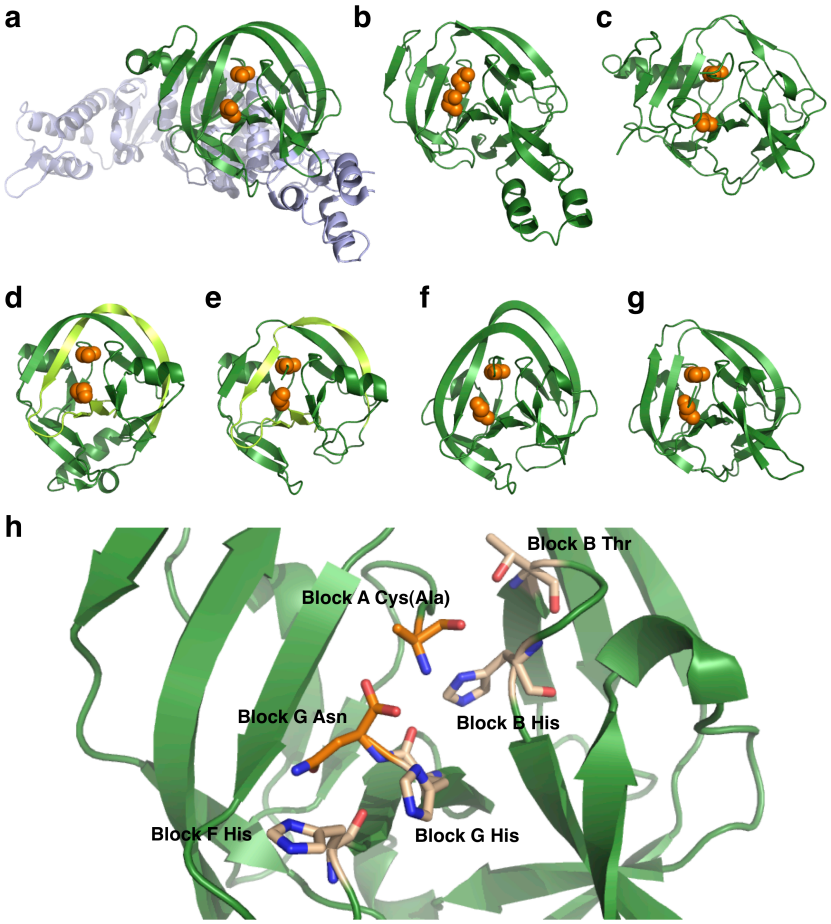


Figure 1.7. Comparison of various Hint domain structures. a. *Thermococcus kodakaraensis* Pol-2 intein (PDB 2CW7).[36] **b.** *Mycobacterium xenopi* GyrA intein (PDB 1AM2).[37] **c.** *Methanococcus jannaschii* KlbA intein (PDB 2JNQ).[27] **d.** *Synechocystis sp.* PCC6803 DnaE split intein (PDB 1ZD7).[38] **e.** *Nostoc punctiforme* DnaE split intein (PDB 2KEQ).[39] **f.** *Clostridium thermocellum* BIL domain 4 (PDB 2LWY).[40] **g.** *Drosophila melanogaster* Hog domain (PDB 1AT0).[35] In panels **a** to **g**, the Hint domain is shown as green ribbon with the Block A nucleophile and Block G asparagine positions highlighted as orange spheres. In panel **a** the homing endonuclease domain in shown in blue. For the split inteins in panels **d** and **e** (which are artificially fused in the solved structures), the C-intein region is light green. **h.** A close up of the *Mxe*GyrA active site, highlighting key residues as sticks. The Block A Cys is mutated to Ala in this structure.

1.3.1. Bacterial intein-like domains

BIL domains can be categorized into two sub-families based on the presence or absence of certain splicing motifs and their resulting biochemical activity.[41] A-type BIL domains have all of the conserved splicing motifs, however most lack the obligatory

cysteine, serine, or threonine at the +1 residue of the C-extein. As a result, they can sustain N- and C-extein cleavage (Figure 1.6). While these BILs should not be splicing-competent, some evidence suggests that they can still splice proteins through an alternate mechanism.[41,42] A small subset of BIL domains actually have a serine or threonine at the +1 position, and these BIL domains can indeed carry out protein splicing in addition to N- and C-extein cleavage. B-type BIL domains have an abnormal Block G that is lacking a C-terminal asparagine but contains a conserved cysteine, serine, or threonine two residues upstream of the canonical intein splice junction. Oddly, these BIL domains are capable of both N- and C-extein cleavage (Figure 1.6), but the latter must proceed through a different mechanism than for inteins and A-type BILs.[41]

While the biochemical difference between BIL domains and inteins is subtle, their phylogenetic distribution differs dramatically. Unlike inteins, which are embedded within highly conserved sites in essential proteins, BILs are integrated into hyper-variable regions of non-conserved proteins.[41] Interestingly, BILs are commonly found attached to secreted proteins, suggesting that they may have a role in protein maturation or translocation. Furthermore, their capacity to activate N-exteins for nucleophilic attack and cleavage without splicing may be utilized for the C-terminal modification of these proteins by any available nucleophile in the cell. Indeed, Nature uses a variation on this mode of post-translational modification in another type of Hint domain.

1.3.2. Hedgehog auto-processing domains

The Hog domain is one of three domains found within the Hedgehog developmental signaling proteins in animals. The N-terminal domain bears the signaling

moiety, the central Hog domain has auto-processing function, and the C-terminal domain binds cholesterol. During the maturation of the Hedgehog signaling proteins, the Hog domain activates the C-terminal amino acid of the signaling region through an N to S acyl shift, analogous to the first step of protein splicing. Following this activation, the cholesterol binding region presents the hydroxyl group of cholesterol as a nucleophile to cleave the signaling domain and tag it with the sterol (Figure 1.8).[43-45] The resulting product is also palmitoylated near its N-terminus through a more traditional mechanism (Figure 1.1a) and secreted from cells to act as a morphogen during developmental patterning.

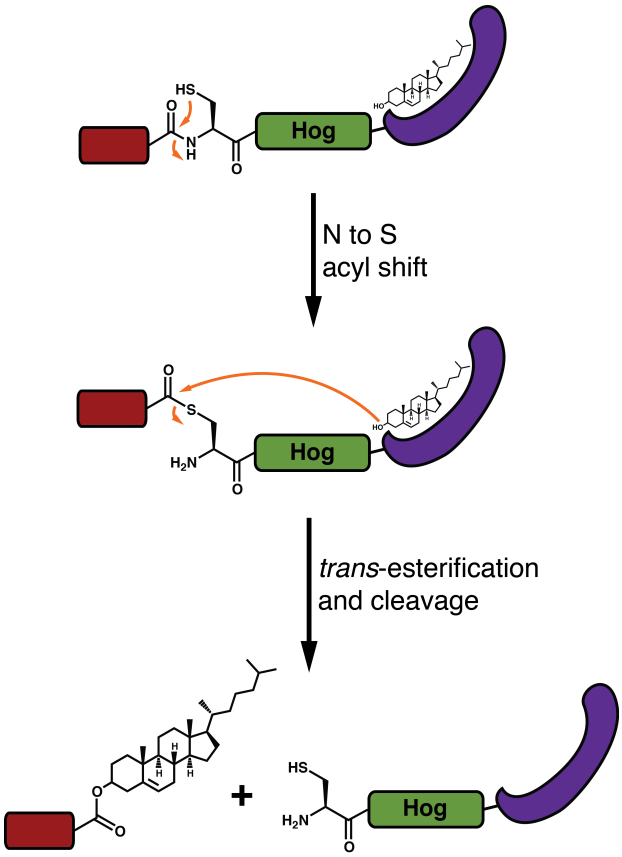


Figure 1.8. Hedgehog auto-processing. The Hog domain (green) activates a peptide bond through an intein-like mechanism, while the choloesterol binding domain (purple) presents the cholesterol hydroxyl group as a nucleophile. The result is N-extein cleavage with concommitant cholesterol ligation.

Hog domains are the only known Hint domains in animals, where they have a clear function during development. The lack of inteins or BIL domains in these higher organisms suggests that a single Hint domain existed in these progenetor organism before the emergence of metazoans, and this ancient Hint domain ultimately became the modern Hog protein involved in development. Consistent with this notion, a secreted protein of unknown function discovered in the genome of the choanoflagellate *Monosiga ovata* was found to be auto-processessed by a Hog domain.[46] As choanoflagellates and metazoans are believed to have diverged from a common ancestor close to the emergence of multicellularity,[47] this protein may provide insights into the evolution of biological function in the Hint superfamily.

1.3.3. The structure of split inteins

Several high-resolution structures of two orthologous cyanobacterial split inteins have been solved. Structures of the DnaE intein from *Synechocystis sp.* PCC6803 (*Ssp*DnaE) were determined using X-ray crystallography (Figure 1.7d),[26,38,48] and structures of the DnaE intein from *Nostoc punctiforme* (*Npu*DnaE) were determined both by nuclear magnetic resonance (NMR) spectroscopy in solution (Figure 1.7e)[39] and by X-ray crystallography.[49] In all cases, the N- and C-intein fragments were recombinantly fused to generate contiguous inteins, and the structures were solved in these fused forms. Regardless, these studies verified that split inteins can adopt the same horseshoe-like fold seen for all other Hint domains. Interestingly, a closer look at the N- and C-intein regions clearly indicates that these fragments are heavily entwined, which would preclude assembly through pre-folded monomers. Furthermore, biophysical measurements indicate

that the *Ssp*DnaE intein fragments associate extremely rapidly (k_{on} of ~10^7 $M^{-1}s^{-1}$) and tightly (K_D of ~30 nM).[50] This highly efficient binding and the entangled topology seen for artificially fused split inteins raise two important questions: (1) Do split inteins retain this topology when they are actually split? (2) If so, how do split intein fragments efficiently arrive at this structure *in trans*? (These issues will be addressed in Chapter 5).

1.4. Applications of protein splicing

In their native context, inteins facilitate both the cleavage and formation of peptide bonds. Early studies on inteins demonstrated that they could also carry out protein splicing in exogenous contexts, often between polypeptides unrelated to the endogenous host protein.[51-53] In fact, the only absolute sequence requirement for intein-mediated splicing outside of the intein itself is the presence of a cysteine, serine, or threonine at the first residue of the C-extein (the +1 position). As such, these molecules should be ideal tools for protein chemistry and engineering. Indeed, intein chemistry is exploited in a variety of powerful technological applications centered around the cleavage and/or formation of peptide bonds.[8]

1.4.1. Tagless protein purification

The side reactions of protein splicing, N- and C-extein cleavage, can both be enhanced by the introduction of specific point mutations in the conserved splicing motifs. For example, mutation of the Block A nucleophile from cysteine/serine to alanine precludes the formation of the linear and branched (thio)ester intermediates. In this context, many inteins have a basal level of asparagine cyclization and thus C-extein

cleavage which can be inhibited at slightly acidic pH.[34,54] Alternatively, mutation of the Block G asparagine to alanine precludes the irreversible branch resolution step, resulting in a trapped equilibrium between the precursor amide and the two (thio)ester intermediates. The (thio)esters can be treated with base to induce N-extein cleavage by hydrolysis.[32,54] These mutant inteins can be used to obtain recombinant proteins without affinity purification tags for use in biochemical studies (Figure 1.9). Specifically, a protein of interest is fused to the N- or C-terminus of an intein bearing the appropriate mutations, and an affinity purification tag is fused to the other intein terminus. After affinity enrichment on a solid support, the pH of the system can be raised to induce N- or C-terminal cleavage, resulting in the release of an untagged protein.[54]

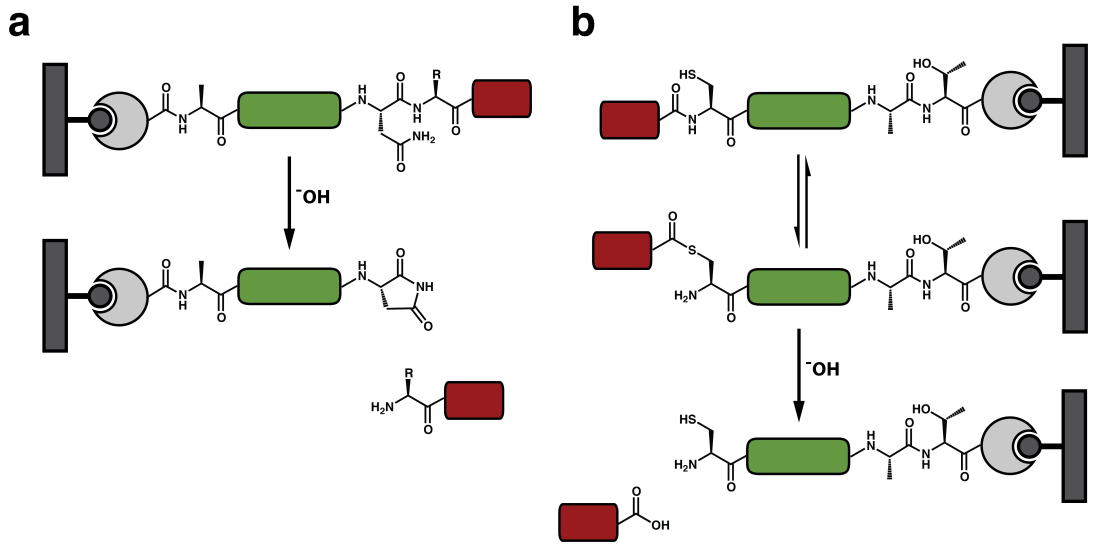


Figure 1.9. Tagless protein purification. a. C-extein cleavage and **b.** N-extein cleavage to yield a tagless affinity-enriched proteins with unmodified N- and C-termini.

1.4.2. In vitro protein semi-synthesis

N-extein cleavage from a mutant intein can be induced not only by hydrolysis but also through attack by nucleophiles other than water.[32] Cleavage using small-molecule alkyl or aryl thiols will result in a protein of interest bearing an α-thioester, rather than a

free carboxylic acid (Figure 1.10a). In the well-known Native Chemical Ligation (NCL) reaction, a peptide C-terminal thioester can be condensed with a 1,2-aminothiol moiety (such as a polypeptide containing an N-terminal cysteine) to form a native peptide bond.[55] While NCL allows for the total chemical synthesis of small proteins through the ligation of synthetic peptides, the use of inteins to generate protein thioesters recombinantly allows for this reaction to be applied to significantly larger macromolecules (Figure 1.10a).[33] This semi-synthetic iteration of NCL, referred to as Expressed Protein Ligation (EPL), has been used to site-specifically incorporate a wide array of chemical moieties into proteins that are typically inaccessible through purely recombinant methods. These include post-translational modifications, unnatural amino acids, backbone modifications, photochemical cross-linkers, biophysical probes, and imaging probes.[8] Finally, it is noteworthy that while intein-mediated EPL traditionally involves the production of the thioester fragment recombinantly, this fragment can also be generated synthetically, and the N-terminal cysteine-containing protein can be made recombinantly, exposing the cysteine either using proteolysis or an intein tag capable of C-extein cleavage.[56]

Protein semi-synthesis by EPL typically employs contiguous inteins that are mutated to prevent protein splicing and generate a reactive handle for fragment condensation. Split inteins offer an alternative approach to protein semi-synthesis, as the complete protein *trans*-splicing (PTS) reaction is effectively a fragment condensation reaction mediated by the complementary N- and C-intein protomers (Figure 1.11b,c). While rates of standard chemical ligation reactions are strongly concentration-dependent, as they rely on the random collision of peptide fragments,[57] PTS by naturally occurring

split inteins is facilitated by a tight protein–protein interaction and thus shows low concentration dependence.[50] This makes split inteins attractive tools for protein semi-synthesis of challenging substrates where sample concentration is a limiting factor.[58]

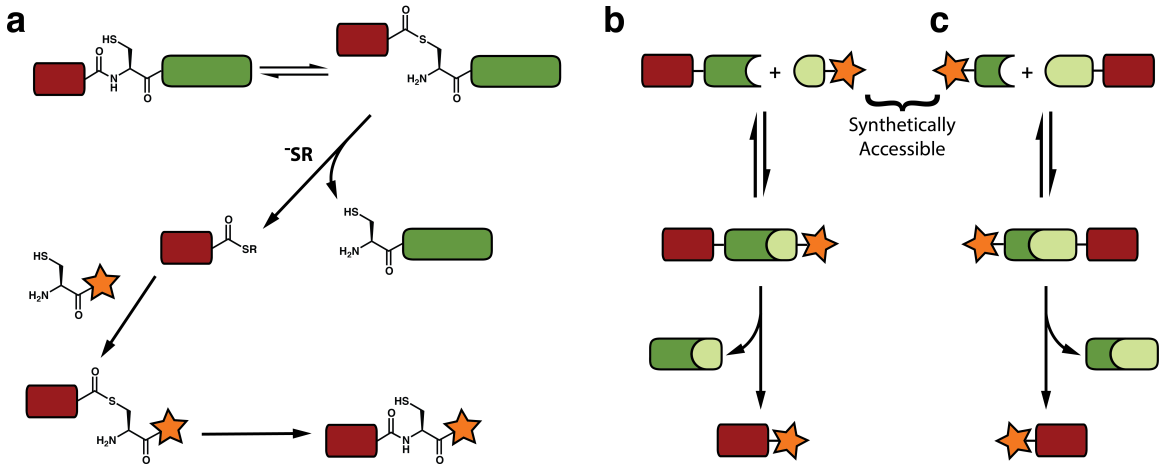


Figure 1.10. ***In vitro*** **protein semi-synthesis. a.** Expressed Protein Ligation (EPL). A C-terminal thioester is generated by cleavage from the intein followed by condensation with an N-terminal cysteine-containing peptide or protein. **b.** Semi-synthesis by protein *trans*-splicing with a synthetically accessible C-intein. **c.** Semi-synthesis by protein *trans*-splicing with a synthetically accessible N-intein.

Importantly, the smaller C-intein fragment of naturally split inteins is typically between 30 and 40 amino acids, making it synthetically accessible through solid-phase peptide synthesis. However the addition of any "cargo" for protein labeling puts this fragment nearly out of synthetic reach. To address this issue, several groups have engineered artificially split inteins with non-canonical split sites as close as 6 residues from the intein C-terminus (Figure 1.10b)[39,59,60] or 11 residues from the N-terminus (Figure 1.10c).[61] While these constructs often have lower splicing efficiencies and weaker binding affinities than naturally split inteins, at least one fragment is synthetically accessible, providing a means to modify the N- or C-terminus of a protein using PTS.[62]

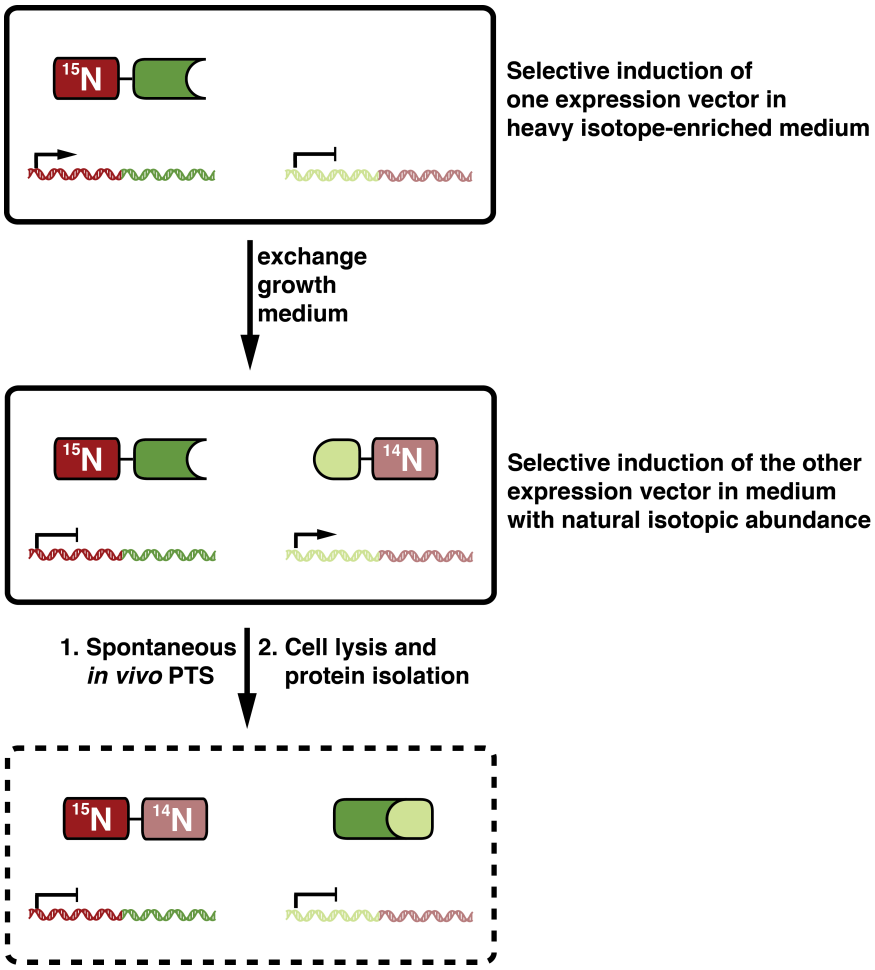


Figure 1.11. Segmental isotopic labeling using split inteins *in vivo*.

1.4.3. Segmental isotopic labeling

Both EPL and PTS are useful intein-based tools for *in vitro* protein semi-synthesis. One of the most important applications of these techniques is the segmental isotopic labeling of proteins for NMR spectroscopy studies. NMR is commonly used to characterize the structure and dynamics of proteins, however this approach requires stable enrichment of NMR-active ^{15}N, ^{13}C, and/or ^{2}H nuclei in the protein of interest, as the natural abundance of these heavy isotopes is too low for practical utility. Standardized techniques now exist to incorporate these isotopes uniformly into the backbone and side chains of recombinant proteins by growing bacterial cells in an isotopically enriched

medium.[63] However, for large proteins (>100 amino acids) and proteins with highly degenerate sequences, significant peak overlap in uniformly labeled samples makes unambiguous interpretation of the spectra challenging.

Using either EPL[64] or PTS,[65] protein fragments with different isotope labeling schemes can be ligated to generate segmentally labeled full-length protein samples with dramatically simplified NMR spectra. Indeed, this approach has been used to aid in the analysis of a wide variety of interesting proteins.[66,67] Two advantages of using PTS with naturally split inteins in this application are that segmentally labeled proteins can be assembled *in vitro* under non-denaturing conditions or *in vivo* through the sequential expression of each fragment in the same cell culture with an intermediate exchange of the medium to include or exclude stable isotope sources (Figure 1.11).[68]

1.4.4. Protein and peptide cyclization

Both EPL and PTS have also been used to generate cyclic polypeptides. Using EPL, a protein can be cyclized when its N-terminus contains an exposed cysteine and its C-terminus is activated through an intein tag to yield an α-thioester. The activated C-terminus can then be directly attacked by the N-terminal cysteine to yield a cyclic thioester, which will rearrange into the more stable peptide bond (Figure 1.12a).[69] The resulting cyclic proteins should have increased thermodynamic and *in vivo* stability, which in turn can improve their biological function, making this an intriguing approach for enhancing the efficacy of protein-based therapeutics.

The efficiency of protein cyclization by EPL is largely dependent on the proximity of the N- and C-termini in the folded state.[70] As an alternative, split inteins can

be used to force the termini together through the association of the intein fragments. To this end, a technique known as *s*plit *i*ntein-mediated *c*ircular *l*igation *o*f *p*eptides and *p*roteins (SICLOPPS) was developed.[71] By inverting the order of intein fragments around a polypeptide of interest, a target sequence can be head-to-tail cyclized (Figure 1.12b). This reaction results in the formation of a native peptide bond upon excision of the N- and C-inteins, leaving behind a single cysteine residue. The power of this technology is two-fold. First, like EPL, it allows for the production of cyclic proteins with enhanced biophysical and biological properties conferred by their augmented stability.[69,72] Perhaps more significantly, it provides a method to generate large libraries of genetically encoded cyclic peptides for rapid screening. Using SICLOPPS, researchers have discovered methyltransferase inhibitors,[73] protease inhibitors,[74] modulators of protein–protein interactions,[75] and molecules that reduce the cellular pathology of Parkinson's disease.[76]

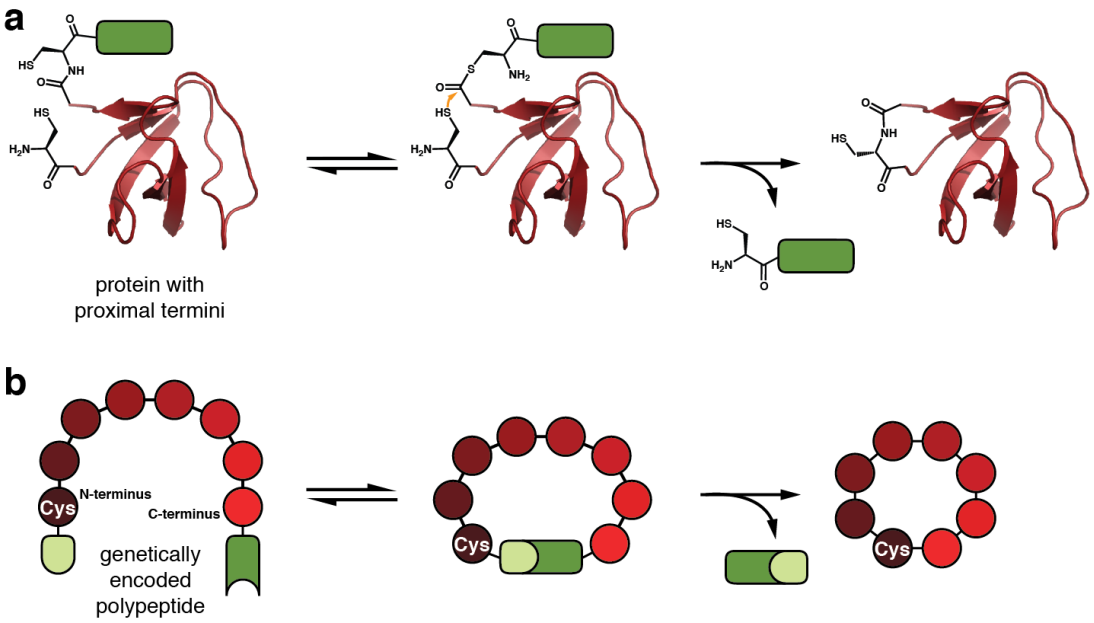


Figure 1.12. Protein and peptide cyclization. a. Cyclization of a protein using EPL. The rendering is based on the N-terminal SH3 domain of c-Crk-II (PDB 1M30), which has been head-to-tail cyclized using this method.[69] **b.** *S*plit *i*ntein-mediated *c*ircular *l*igation *o*f *p*eptides and *p*roteins (SICLOPPS).

1.4.5. Conditional protein splicing

Conditional protein splicing (CPS), the activation or inhibition of protein splicing by an extrinsic modulator, is perhaps the most intriguing application of inteins. The basic premise of CPS is that inteins control the primary sequence, and thus function, of the proteins they are splicing, so controlling intein activity would provide a way to "turn on" any protein at will, even *in vivo*. Current CPS systems come in three flavors: (1) Contiguous inteins have been fused to ligand binding domains that can allosterically modulate protein splicing in response to a small molecule (Figure 1.13a).[77,78] (2) Both contiguous and naturally split inteins have been chemically caged at or near active-site residues to control splicing in response to light or proteolysis (Figure 1.13b).[79,80] (3) Artificially split inteins, which cannot spontaneously associate, have been fused to heterodimerization domains to reconstitute their splicing activity in response to small molecules or light (Figure 1.13c).[81-83]

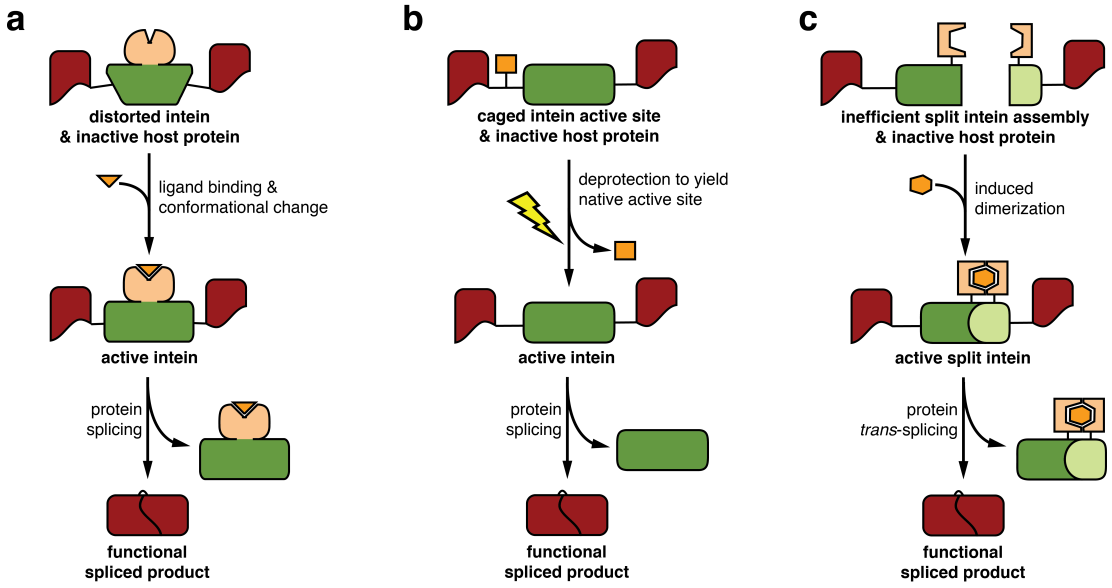


Figure 1.13. Conditional protein splicing. a. Intein activation through conformational change induced by ligand binding to a fused ligand binding domain. **b.** Intein activation through deprotection of a photo-caged active site residue. **c.** Activation of an artificially split intein through chemically induced dimerization.

The post-translational activation of protein function in response to these exogenous factors should be faster than traditional molecular biology techniques involving inducible promoters, tunable in a dose-dependent manner, and portable, as inteins can splice in a wide variety of protein contexts. Given these facts, the CPS systems described above are promising tools for cell biology and for the development of "smart" protein therapeutics that are only activated at the appropriate site of action.

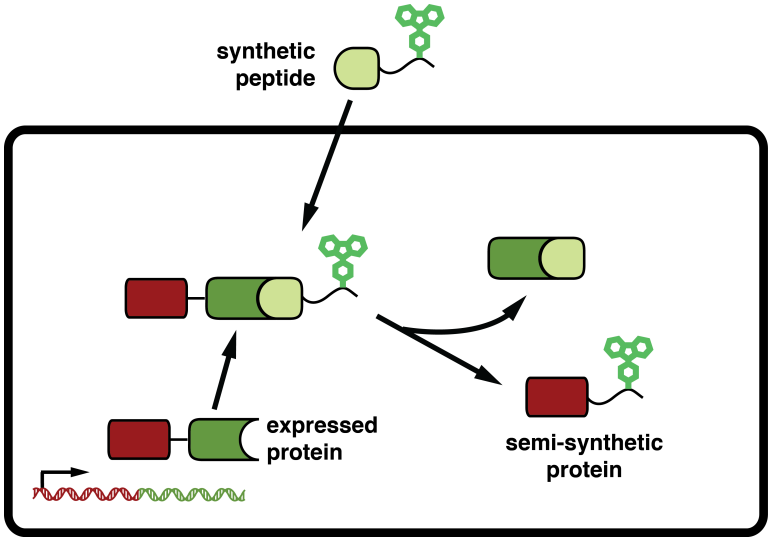


Figure 1.14. *In vivo* protein semi-synthesis. A fluorophore is synthetically attached to the C-intein, and the construct is delivered into live cells. A protein of interest is expressed as a fusion to an N-intein. Upon delivery of the C-intein and its cargo, protein *trans*-splicing spontaneously occurs, generating a semi-synthetic protein in the cell.

1.4.6. In vivo protein semi-synthesis

The most significant advantage of PTS-based protein semi-synthesis over EPL with contiguous inteins is that it can be readily applied *in vivo*. In purified systems, reaction specificity is governed simply by the functional groups present in the molecules of interest, but living systems are heterogeneous and chemically complex, making orthogonal chemistry more challenging.[84] Split inteins overcome this problem by acting as ligation auxiliaries that engender virtually absolute specificity to the ligation reaction

of interest. Indeed PTS has been used *in vivo* to site-specifically label proteins with a synthetic probe[85,86] and to construct a large gene product from its fragments.[87] In the former studies, a fluorophore was attached to the synethically accessible C-intein, and, using a protein transduction system, the construct was readily delivered into cells expressing the N-intein fused to a target protein of interest. The intein fragments readily associated and spliced, thereby site-specifically labeling a protein in living cells (Figure 1.14).[85,86] In the later study, the authors overcame the size-limitations of adeno-associated viral vectors for gene therapy by splitting the cargo into two fragments and assembling the therapeutic target *in vivo* by PTS.[87] Importantly, this work demonstrates the potential utility of split inteins as more than just a research tool.

1.5. Caveats of protein splicing

While numerous intein-based technologies have been developed and widely-used, the full scope of their utility is limited by two common properties of most inteins: (1) slow splicing and cleavage reactions and (2) dependence on local extein sequence composition.

1.5.1. Reaction kinetics and yield

Most commonly used inteins carry out splicing reactions slowly, on the order of hours. For example, the contiguous *Mxe*GyrA intein, frequently used to generate protein thioesters for EPL, carries out the overall splicing reaction with a half-life of several hours at 25 °C.[30] The first-discovered split intein, *Ssp*DnaE intein, has a half-life of three hours at 23 °C.[88] Furthermore, at 37 °C, the Ssp intein has an increased rate of N-extein

cleavage, lowering the overall yield of the splicing reaction.[89] Recently, however, two groups demonstrated that the related *Npu*DnaE split intein could carry out protein *trans*-splicing with extremely high yields *in vivo*[90] and with a half-life of 63 seconds *in vitro*,[89] both at 37 °C. This anomalous splicing efficiency prompts the intriguing questions: "Is the Npu intein an outlier?" and "How can this intein splice so rapidly?"

1.5.2. Extein dependence

Even the Npu intein, with its remarkable splicing kinetics, suffers from another pervasive functional caveat of all inteins: its reaction kinetics and overall yield are highly dependent on the identity of extein residues surrounding the splice junctions (Figure 1.5a, positions -3, -2, -1, +2, and +3). Typically, inteins have the fastest reaction rates and least side reactions when embedded between the local extein residues found in their endogenous host protein. Deviation at the local N-extein residues (especially the -1 position) can have a profound effect on the initial N to O/S acyl shift reaction,[54] which has practical implications for thioester formation during EPL. Deviation at the local C-extein residues (especially the +2 C-extein position) has been shown to dramatically reduce the overall splicing rate and yield for the Npu and Ssp split inteins,[90-92] thus requiring that native extein residues be added to a target protein during protein semi-synthesis and segmental isotope labeling using *trans*-splicing.[93]

1.5.3. Improving intein-based technologies

While the aforementioned caveats apparently call into question the broad utility of inteins for protein chemistry and engineering, the gradual increase in intein technology-

related papers published since the discovery of protein splicing argues that these shortcomings have not inhibited practical intein use (Figure 1.15). Furthermore, substantial efforts have been made to overcome the deficiencies of currently used inteins. For example, several groups have applied directed evolution techniques to improve splicing and cleavage rates,[94] pH sensitivity,[34] tolerance to non-native extein residues,[92,95] temperature dependence,[92] and even conditionality in response to a small molecule.[78]

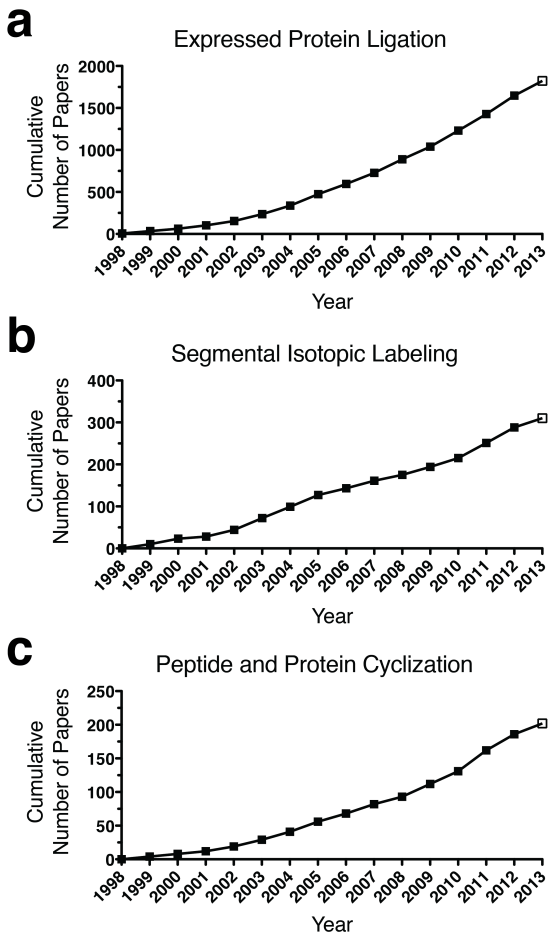


Figure 1.15. Intein technology-related publications. The cumulative number of publications up to each year is based on Google Scholar searches for the following topics: **a.** "expressed protein ligation", **b.** "segmental isotopic labeling" or "segmental isotope labeling" **c.** "protein cyclization" AND "intein" or "peptide cyclization" AND "intein". Note that these statistics are meant to be a rough indication of papers published each year utilizing those intein-based technologies. Some counted publications may not utilize the searched technique but merely mention it, while other relevant publications may not have emerged under these search criteria. The open box for 2013 signifies that this number only reflects data up to October of that year.

In addition, the rigorous characterization of intein structure and reactivity has opened avenues for the rational design of more useful inteins. Indeed, elucidation of the protein splicing mechanism guided researchers towards the first EPL platforms,[32,33] and an understanding of intein secondary structure topology and backbone dynamics has aided in the identification of non-canonical split sites to make synthetically accessible split intein fragments.[39,59-61] These successes strongly indicate that basic intein research can fuel the development and improvement of useful intein based technologies.

1.6. Summary and conclusions

Inteins are a curious class of auto-processing domains that are adept at breaking and making peptide bonds. Their ancient origins and pervasiveness in nature suggest that these proteins once had an important biological function. While only remnants of this natural function are evident now in intein-related proteins, the utility of inteins in a technological setting is clear. These proteins have been used in countless ways as protein engineering tools for both basic and applied research.

Naturally split inteins are particularly intriguing, both from a basic biochemical/biophysical standpoint, as well as for their applications. Their capacity to carry out protein spicing *in trans* can provide additional benefits when applied to each of the techniques discussed above. Furthermore, the discovery of the fast-splicing split DnaE intein from *Nostoc punctiforme* (*Npu*DnaE or Npu) should vastly improve intein-based technologies and make them more applicable to biological systems.

The unprecedented splicing kinetics for the Npu intein were first reported in February 2009, one month before I began to explore the intein field.[89] As such, this

discovery inspired my research, and this remarkable protein is at the core of each chapter in this thesis. In Chapter 2, the apparent uniqueness of Npu is explored by comparing the reactivities of several orthologous split inteins. These experiments are used to understand the molecular determinants for rapid protein splicing and to develop a streamlined version of Expressed Protein Ligation. In Chapter 3, the issue of extein-dependent splicing kinetics in Npu is rigorously addressed using a combination of protein semi-synthesis, kinetic analysis, and structural analysis. A clear structural role for specific extein residues in the protein splicing reaction is established, suggesting that there may be a way to alleviate extein-dependence through protein engineering. In Chapter 4, a property unique to split inteins, charge-segregation between the N- and C-terminal fragments, is analyzed in the context of Npu fragment association and reactivity. These experiments provide the first insight into what drives the assembly of an intein domain *in trans* and allow for the development of new protein semi-synthesis tools. In Chapter 5, the detailed mechanism of Npu fragment assembly is elucidated using a combination of protein semi-synthesis and biophysical techniques. This mechanism of coupled binding and folding has both evolutionary and technological implications for split inteins.

The work described in this thesis is guided by a single unifying tenet: that basic research to understand the biochemical and biophysical properties of inteins will ultimately enhance their utility in a practical setting. Following this theme, in Chapter 6, future prospects for intein research are discussed. First, building on the work presented in the previous chapters, approaches to solve persistent issues with protein *trans*-splicing technologies are explored. Specifically the issues of extein dependence, synthetic inaccessibility of native split intein fragments, and the fact that fragment assembly of

naturally split inteins cannot be controlled are discussed. Finally, open questions in basic intein research are addressed, including the elusive (and possibly non-existent) biological role of inteins in their natural context.

Chapter 2: Ultrafast split DnaE inteins and their use in protein engineering*

Given their capacity to make and break peptide bonds, inteins have found widespread use as chemical biological tools.[8] Despite the growing use of inteins in chemical biology, however, their practical utility has been constrained by two common characteristics, namely (1) slow kinetics[30,88,89] and (2) context dependent efficiency with respect to the immediate flanking extein sequences.[54,91] Recently, a split intein from the cyanobacterium *Nostoc punctiforme* (Npu) was shown to catalyze protein *trans*-splicing on the order of a minute, rather than hours like most *cis*- or *trans*-splicing inteins.[89] Furthermore, this intein was slightly more tolerant of sequence variation on the proximal C-extein residues than its slow-splicing ortholog from *Synechocystis sp.* PCC6803 (Ssp).[90]

We became interested in the apparently unique properties of Npu and sought to determine whether other homologous split inteins could also catalyze rapid *trans*-splicing, perhaps with greater C-extein tolerance. Of the roughly 600 inteins currently catalogued,[10] less than 5% are split inteins, mostly from a family known as the cyanobacterial split DnaE inteins (Table 2.1 and Figure 2.1).[19] Surprisingly, only six of these, including Npu, have been experimentally analyzed to any extent,[90,96,97] and only the kinetics of Npu and its widely-studied ortholog, Ssp, have been rigorously

---

* The work in this chapter was carried out in close collaboration with Miquel Vila-Perelló and Zhihua Liu and was published in the following two papers:

Shah, N. H., Dann, G. P., Vila-Perelló, M., Liu, Z., and Muir, T. W. Ultrafast protein splicing is common among cyanobacterial split inteins: Implications for protein engineering. *J. Am. Chem. Soc.* **134**, 11338-11341 (2012).

Vila-Perelló, M., Liu, Z., Shah, N. H., Willis, J. A., Idoyaga, J., and Muir, T.W. Streamlined expressed protein ligation using split inteins. *J. Am. Chem. Soc.* **135**, 286-292 (2013).

characterized *in vitro*.[88,89] We hypothesized that thorough a functional characterization of this entire family would not only uncover new efficient inteins but also provide insights into the specific features of these inteins that allow them to sustain high activity.

Table 2.1. Split DnaE intein names and host organisms.

DnaE Intein Name	Genus	Species	Strain
Npu	*Nostoc*	*punctiforme*	PCC73102
Ssp	*Synechocystis*	*species*	PCC6803
Aha	*Aphanothece*	*halophytica*	
Aov	*Aphanizomenon*	*ovalisporum*	
Asp	*Anabaena*	species	PCC7120
Ava	*Anabaena*	*variabilis*	ATCC29413
Cra(CS505)	*Cylindrospermopsis*	*raciborskii*	CS-505
Csp(CCY0110)	*Cyanothece*	species	CCY0110
Csp(PCC8801)	*Cyanothece*	species	PCC8801
Cwa	*Crocosphaera*	*watsonii*	WH 8501
Maer(NIES843)	*Microcystis*	*aeruginosa*	NIES-843
Mcht(PCC7420)-2	*Microcoleus*	*chthonoplastes*	PCC7420
Oli	*Oscillatoria*	*limnetica*	Solar Lake
Sel(PC7942)	*Synechococcus*	*elongatus*	PC7942
Ssp(PCC7002)	*Synechococcus*	species	PCC7002
Tel	*Thermosynechococcus*	*elongatus*	BP-1
Ter-3	*Trichodesmium*	*erythraeum*	IMS101
Tvu	*Thermosynechococcus*	*vulcanus*	

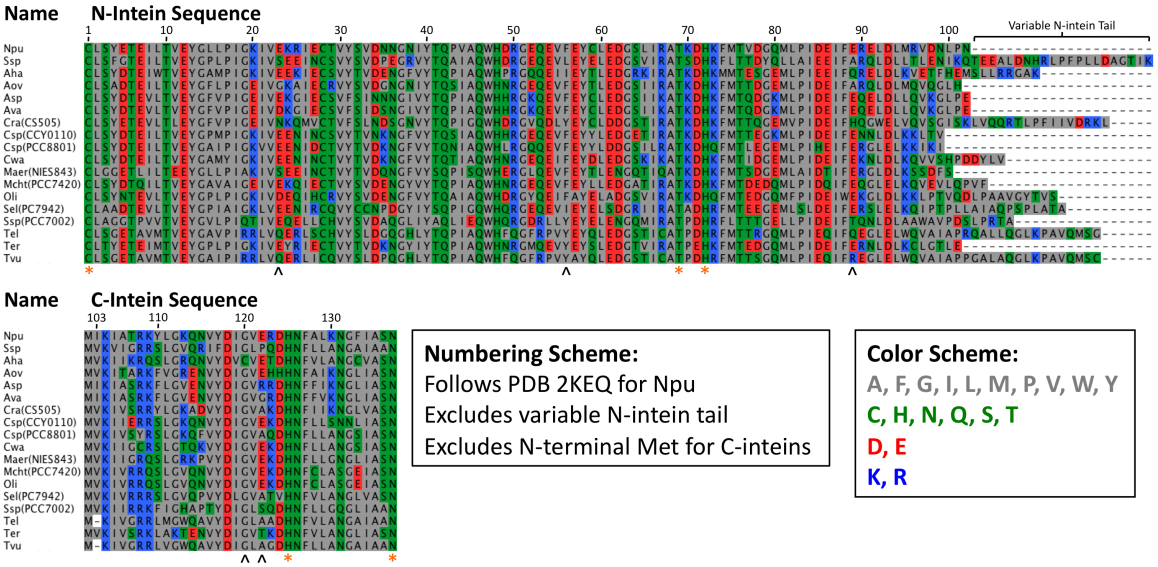


Figure 2.1. Sequence alignment of split DnaE inteins. Numbering follows that of Npu as assigned for the NMR structure (PDB 2KEQ).[39] The N-terminal methionine of the C-intein is not included in the numbering. Critical catalytic residues are marked with an orange asterisk and positions mutated in the sequence-activity correlation studies discussed in the text are marked with a black carrot.

2.1. Parallel characterization of 18 split inteins *in vivo*

We began our investigation with a rapid survey of 18 split DnaE inteins. Previously, we described an *in vivo* screening method to accurately compare the efficiencies of split inteins.[92,98] In this assay, the two fragments of a split intein are co-expressed in *E. coli* as fusions to a fragmented aminoglycoside phosphotransferase (KanR) enzyme. Upon *trans*-splicing, the functional enzyme is assembled, and the bacteria become resistant to the antibiotic kanamycin (Figure 2.2a). More active inteins confer greater kanamycin resistance and thus have a higher IC_{50} value for bacterial growth as a function of kanamycin concentration (Figure 2.2b,c).

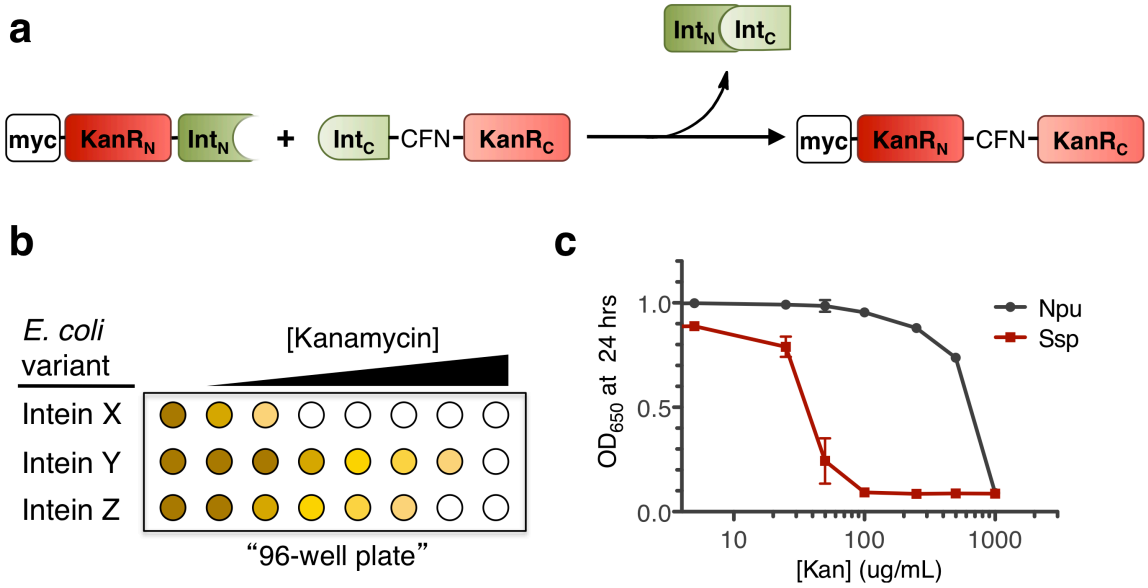


Figure 2.2. *In vivo* screening assay for split intein activity. a. Scheme depicting intein activity-dependent kanamycin resistance. **b.** Scheme depicting multi-well format screening approach. **c.** Kanamycin dose-response curves for Npu and Ssp *in vivo* splicing.

Since all DnaE inteins splice the same local extein sequences in their endogenous context,[19] we carried out our screen in a wild-type C-extein background (CFN) within the KanR enzyme (Figure 2.2a). As expected, bacteria expressing the Npu intein had a high IC_{50}, whereas clones expressing Ssp showed poor resistance to kanamycin (Figure 2.2c).

Remarkably, more than half of the DnaE inteins showed splicing efficiency comparable to Npu *in vivo* at 30°C (Figure 2.3).

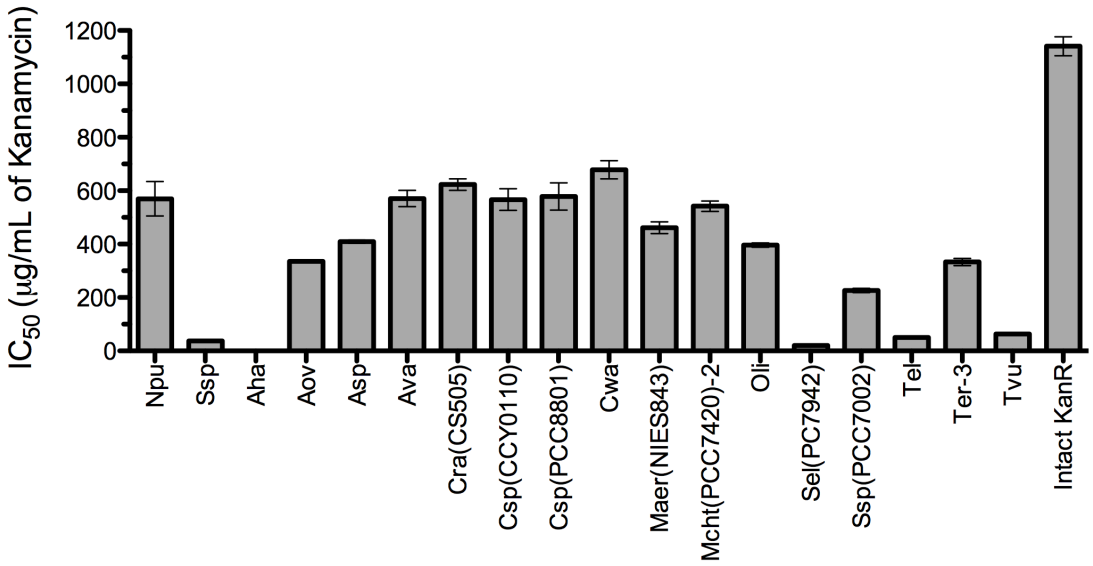


Figure 2.3. *Trans*-splicing of split DnaE inteins *in vivo*. Relative *trans*-splicing efficiencies at 30°C were determined using the assay depicted in Figure 2.2. IC_{50} values were extracted from dose-response data like that shown in Figure 2.2c (± SE, n = 3-4).

2.2. *In vitro* validation of conserved "ultrafast" protein *trans*-splicing

To confirm that the high IC_{50} values observed *in vivo* reflected rapid *trans*-splicing, we performed a series of kinetic studies under standardized conditions *in vitro*. For this, we individually expressed and purified several of the split DnaE intein fragments fused to model N- and C-extein domains, ubiquitin and SUMO, respectively. Importantly, we preserved the endogenous local extein residues as linkers between the extein domains and intein fragments to recapitulate a wild-type-like splicing context (Figure 2.4a). Cognate intein fragments were mixed at 1 μM, and the formation of the Ub-SUMO spliced product at 30°C and 37°C was monitored by gel electrophoresis (Figure 2.4b,c). These assays validated that the new inteins with high-activity *in vivo* could catalyze *trans*-splicing *in vitro* in tens of seconds, substantially faster than Ssp (Figure 2.4d,e and Table 2.2). Interestingly, all of the inteins analyzed except Ssp showed

increased splicing rates at 37°C. Furthermore, all of the fast-splicing inteins showed low-to-undetectable levels of side reactions, again in contrast to Ssp, which had marked N-extein and C-extein cleavage (Figure 2.4b,c).

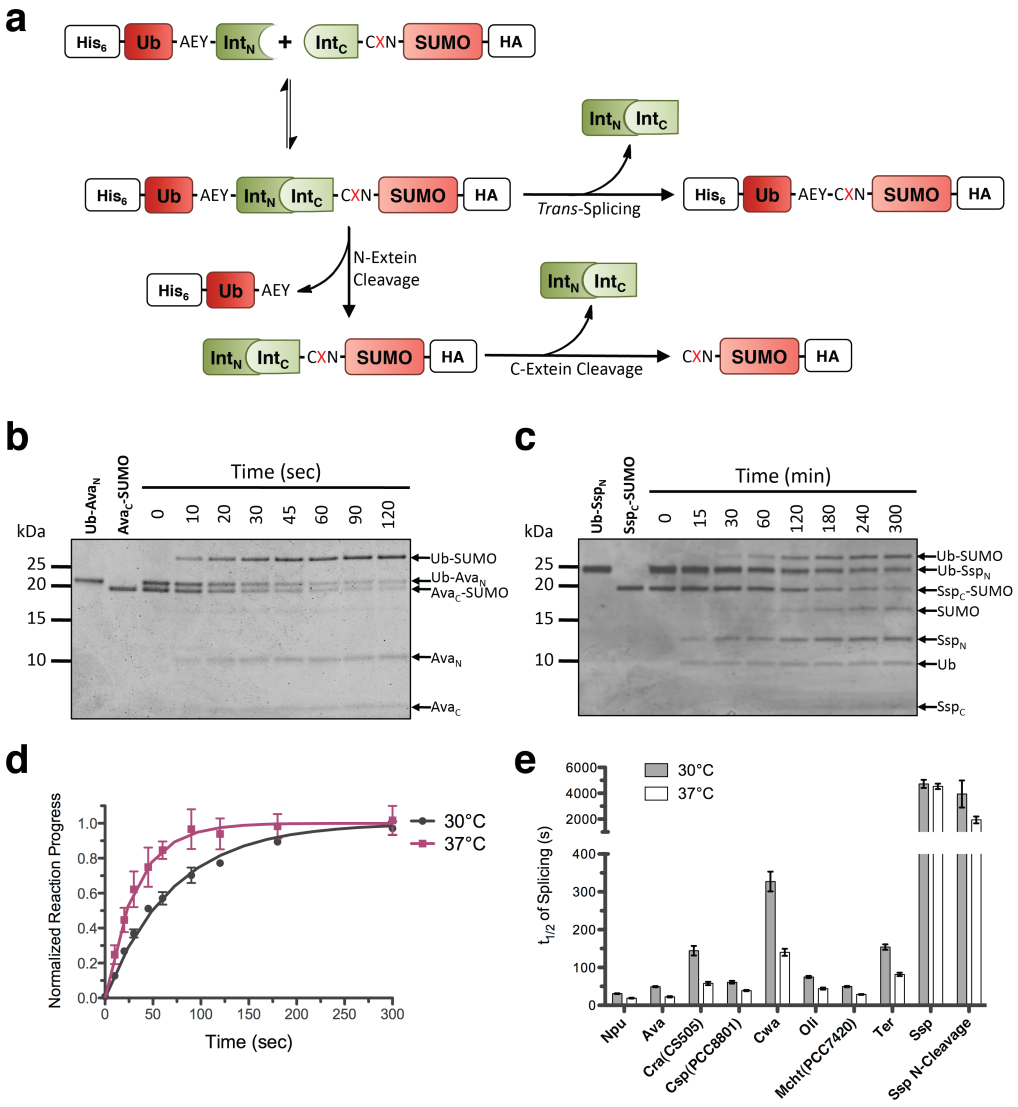


Figure 2.4. *In vitro trans*-splicing assays. a. Scheme depicting the *in vitro* splicing assay, showing possible products and side products. C-extein +2 residue "X" was the native phenylalanine for the assays described in this chapter. Coomassie-stained SDS-PAGE gels showing **b.** fast splicing of the Ava split intein and **c.** inefficient splicing of the Ssp split intein at 37 °C. **d.** Reaction progress curves extracted from densitometric analysis of Ava splicing assays (± SD, n = 3). **e.** Half-lives extracted from the reaction progress curves fit to a first-order rate equation (± SE, n = 3). Reactions were carried out in a pH 7.2 buffer containing 100 mM phosphates, 150 mM NaCl, 1 mM DTT, and 1mM EDTA and time points were quenched in SDS gel loading dye containing β-mercaptoethanol.

Table 2.2. Observed first-order rate constants for *in vitro* splicing reactions.[a]

Intein	k_{obs} (s^{-1}), 30°C	k_{obs} (s^{-1}), 37°C
Npu	(2.2 ± 0.8) x 10^{-2}	(3.7 ± 0.2) x 10^{-2}
Ava	(1.4 ± 0.4) x 10^{-2}	(3.1 ± 0.2) x 10^{-2}
Cra(CS505)	(4.8 ± 0.4) x 10^{-3}	(1.2 ± 0.1) x 10^{-2}
Csp(PCC8801)	(1.1 ± 0.1) x 10^{-2}	(1.8 ± 0.1) x 10^{-2}
Cwa	(2.1 ± 0.2) x 10^{-3}	(5.0 ± 0.3) x 10^{-3}
Oli	(9.2 ± 0.3) x 10^{-3}	(1.6 ± 0.1) x 10^{-2}
Mcht(PCC7420)	(1.4 ± 0.1) x 10^{-2}	(2.4 ± 0.1) x 10^{-2}
Ter	(4.5 ± 0.2) x 10^{-3}	(8.5 ± 0.5) x 10^{-3}
Ssp	(1.5 ± 0.1) x 10^{-4}	(1.5 ± 0.1) x 10^{-4}
Ssp, Ub-Cleavage	(1.8 ± 0.5) x 10^{-4}	(3.6 ± 0.5) x 10^{-4}

[a] The indicated error is the standard error in the best-fit rate constant from a standard first-order rate equation, n = 3.

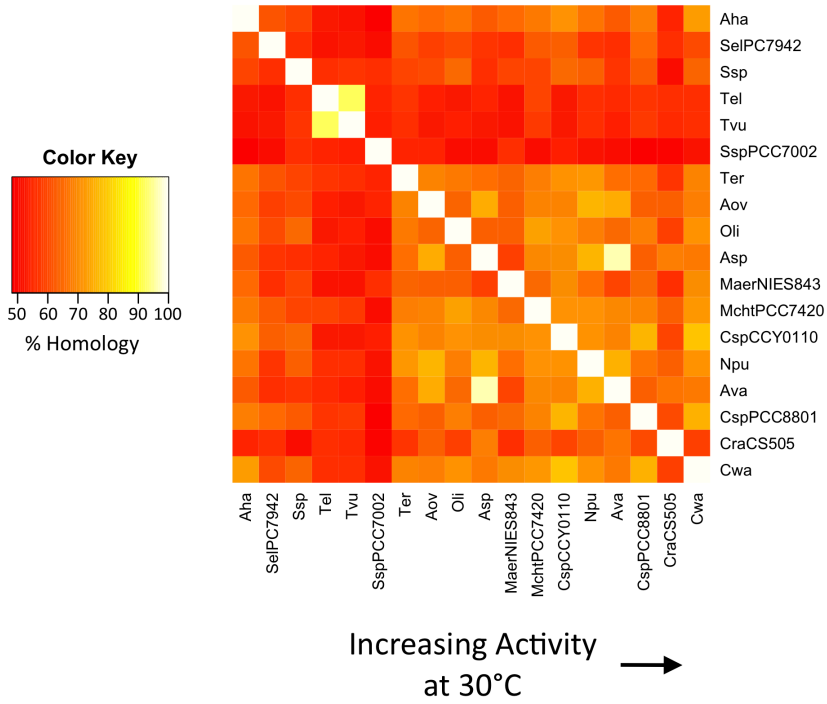


Figure 2.5. Correlation between global sequence identity and *in vivo* activity. Percent sequence identity is calculated based on the alignment shown in Figure 2.1, excluding the variable N-intein tail.

2.3. Sequence-activity correlations in the DnaE inteins

Our data indicate that the split DnaE inteins are highly divergent in activity, despite all having evolved to catalyze *trans*-splicing on virtually identical substrates (the highly conserved cyanobacterial DnaE protein). Interestingly, the key catalytic residues

involved in splicing are conserved across the entire family (Figure 2.1, orange asterisks). Thus, residues that affect splicing activity are non-catalytic and perhaps only moderately conserved. We envisioned that our measurements of relative activity could facilitate the discovery of specific sequence features that differentiate high-activity inteins from inefficient ones. Indeed, sequence homology analysis indicates that inteins with high activity are more homologous to one another than they are to the low-activity inteins (Figure 2.5). One significant outlier to this observation is the intein from *Aphanothece halophytica* (Aha) which, despite having greater than 65% sequence identity to the high-activity inteins, had no detectable activity *in vivo*. Closer inspection of a multiple sequence alignment indicated that this intein has a non-catalytic cysteine (position 120) in place of an otherwise absolutely conserved glycine (Figure 2.1). Furthermore, this position is close to the intein active site (Figure 2.6a), where an extra nucleophile may facilitate N-extein cleavage. Gratifyingly, mutating this cysteine to glycine almost completely suppressed N-extein cleavage and reinstated high activity in the Aha intein (Figure 2.6b,c), validating the predictive capacity of our data.

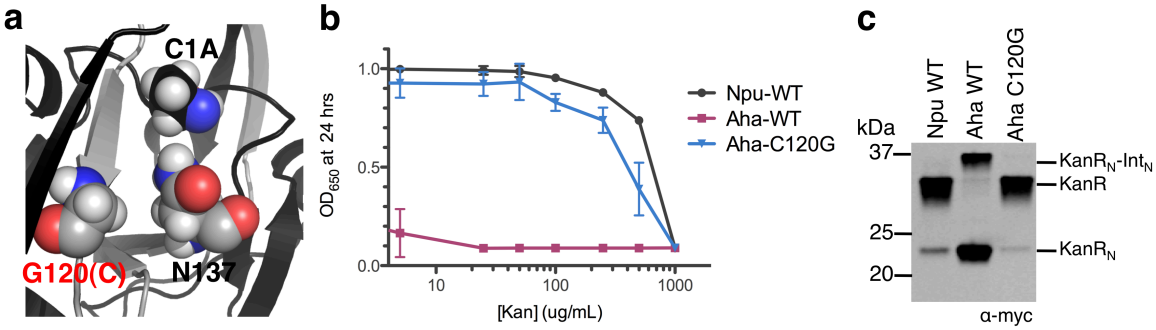


Figure 2.6. Reinstating predicted activity in the Aha intein. a. Rendering of the Npu structure[39] highlighting the proximity of position 120 to the terminal catalytic residues C1 and N137. **b.** Kanamycin dose-response curves for *in vivo* splicing of the Aha C120G mutation (± SD, n = 3). **c.** Western blot analysis of whole cell lysates from cells expressing wild-type Npu, wild-type Aha, or Aha(C120G). A myc-tag is present at the N-terminus of the $KanR_N$ protein (Figure 2.2a).

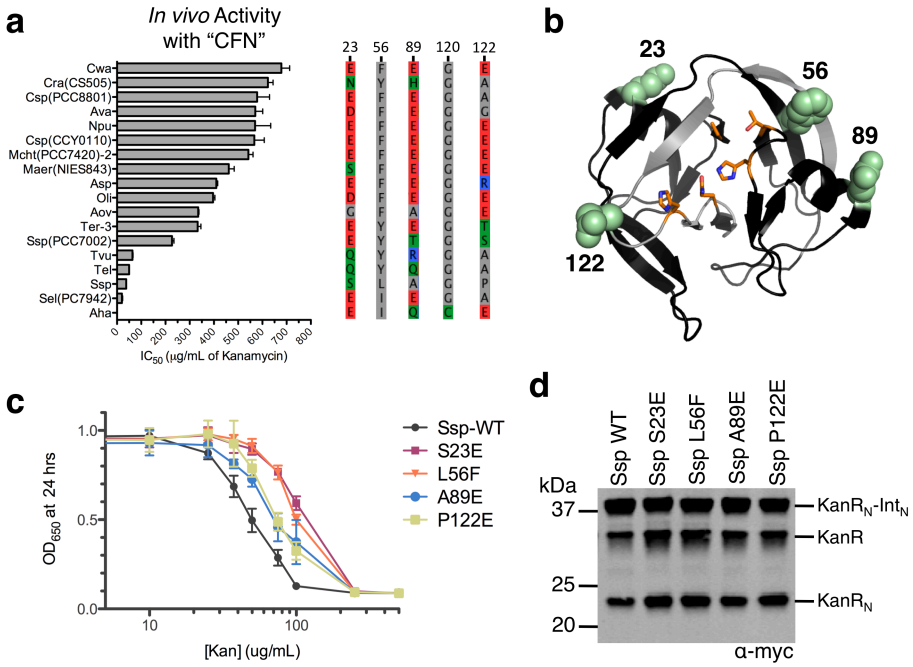


Figure 2.7. Sequence-activity relationships in split DnaE inteins. a. Inteins in order of *in vivo* splicing activity with selected slices from the corresponding multiple sequence alignment. **b.** Rendering of the Npu structure[39] highlighting key catalytic residues (orange sticks) and important non-catalytic positions (green spheres) that modulate Ssp activity. **c.** Kanamycin dose-response curves for *in vivo* analysis of Ssp-to-Npu point mutations that improve Ssp activity (± SD, n = 4). **d.** Western blot analysis of whole cell lysates from cells expressing wild-type Ssp and four point mutants. Note, all residue numberings correspond to the relevant positions on Npu as defined in Figure 2.1.

Further analysis of the split intein sequence alignment indicated that several positions have strong amino acid conservation amongst the high-activity inteins but diverge for the low-activity inteins. These may be sites where the fast inteins have retained beneficial interactions that have been lost in slow ones. To test this idea, we chose several positions where this sequence-activity correlation was apparent and replaced the residue in Ssp with the corresponding amino acid found in the fast inteins (Figure 2.7a,b). Consistent with our hypothesis, several point mutations increased the activity of Ssp *in vivo* (Figure 2.7c,d). While the specific roles of these residues are not explicitly clear, especially given that they lie outside of the active site (Figure 2.7b), their

locations on the intein fold[39] may provide some insights into their function. For example, at position 56, an aromatic residue is preferred in the high-activity inteins. This position is adjacent to the conserved catalytic TXXH motif (positions 69-72), and an aromatic residue may facilitate packing interactions to stabilize those residues. Similarly, a glutamate is preferred at position 122, proximal to catalytic histidine 125 (discussed further in Chapter 3). The glutamate at position 89 is involved in an intimate ion cluster that is important for stabilizing the split intein complex (discussed further in Chapters 4 and 5).[98] Interestingly, E23 is distant from the catalytic site and has no obvious structural role. This position is conceivably important for fold stability or dynamics as has previously been observed for distal activating point mutations in other inteins.[95,99]

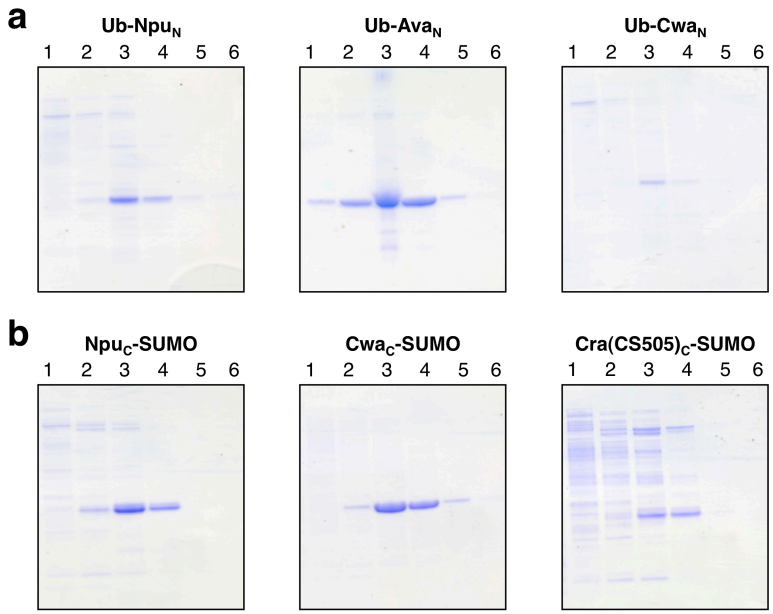


Figure 2.8. SDS-PAGE analysis of Ni-enriched Ub-Int$_N$ and Int$_C$-SUMO fusions. a. Ni column fractions for three Ub-Int$_N$ constructs. **b.** Ni column fractions for three Int$_C$ constructs. All proteins were expressed and enriched identically. The six lanes in each gel correspond to (1) a 5 column volume (CV) wash with 20 mM imidazole, (2) a 3 CV wash with 50 mM imidizole, (3)-(6) four 1.5 CV elutions with 250 mM imidazole. Note that the smearing observed for Ub-AvaN is a result of extremely high protein yield and consequent gel overloading. Gels were coomassie-stained for visualization. Proteins were further purified by size exclusion prior to the activity assays described above.

2.4. Practical considerations with new fast split inteins

The discovery of new, fast *trans*-splicing inteins has broad implications for protein chemistry. Indeed, the discovery of Npu fueled a resurgence in the use of split intein-based technologies.[98,100,101] While no single intein may be ideal for every protein chemistry endeavor, the availability of several new fast-splicing split inteins should provide options to enhance the efficiency of most *trans*-splicing applications. For example, one common problem in working with split inteins is low expression yield or poor solubility of an intein fragment fusion to a protein of interest. Indeed, our over-expression and purification efforts showed that the Ub-Int_N and Int_C-SUMO fusions have markedly different yields of soluble expression, depending on the intein (Figure 2.8). Thus, a short-list of highly active split inteins with varying behavior will serve as a starting point for empirical optimization of a given *trans*-splicing application. Furthermore, the fragments of the different fast-splicing split inteins can be mixed as non-cognate pairs and still retain highly efficient splicing activity, further expanding the options available for any *trans*-splicing application (Figure 2.9).

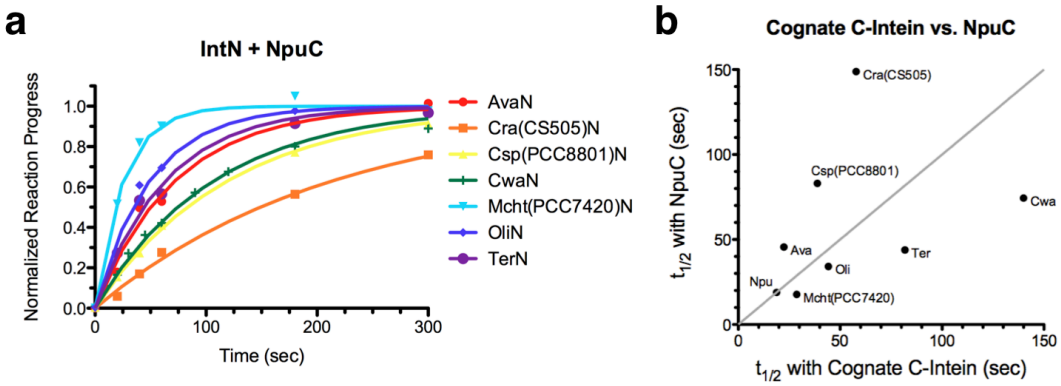
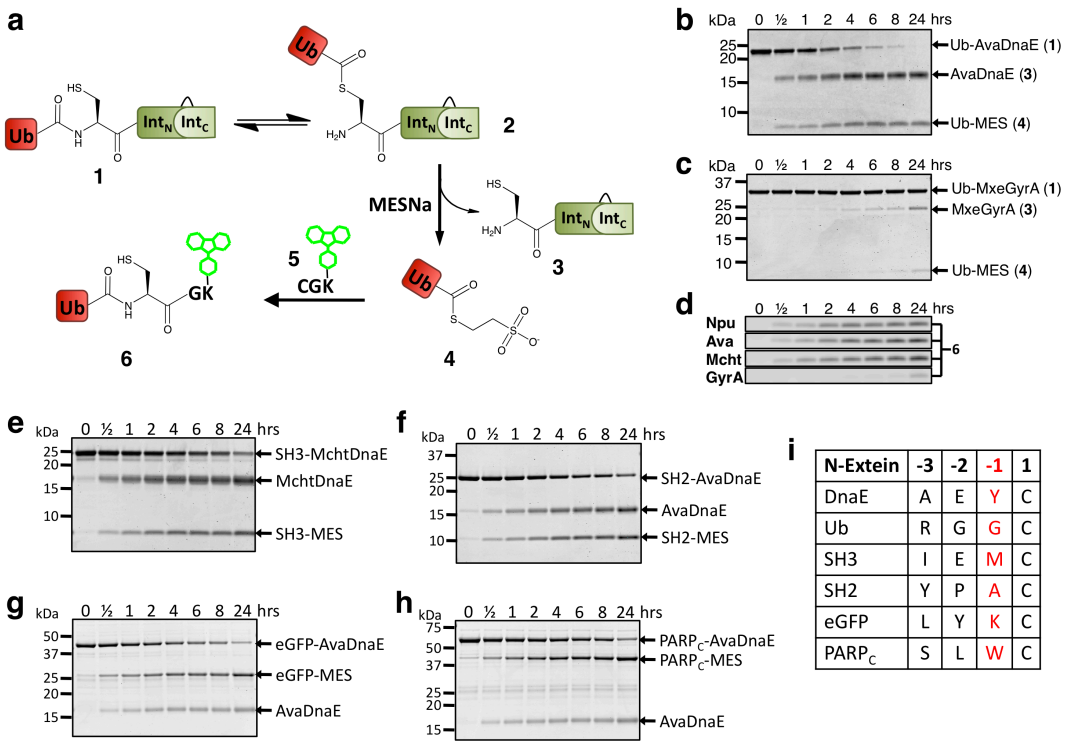


Figure 2.9. Cross-reactivity of N-inteins with Npu_C *in vitro*. a. Reaction progress curve for 7 non-cognate N-inteins reacting with Npu_C in the *in vitro* gel-base splicing assay. **b.** Comparison of half-lives, extracted from first-order fits of the curves in **a**, for N-inteins with their cognate C-intein versus with Npu_C.



N-Extein	-3	-2	-1	1
DnaE	A	E	Y	C
Ub	R	G	G	C
SH3	I	E	M	C
SH2	Y	P	A	C
eGFP	L	Y	K	C
PARP$_C$	S	L	W	C

Figure 2.10. Efficient EPL using contiguous DnaE inteins. a. Scheme showing the formation of the linear thioester intermediate and its use to generate a protein α-thioester for EPL. Coomassie-stained SDS-PAGE gels depicting MESNa thiolysis of ubiquitin from **b.** the fused AvaDnaE intein and **c.** the MxeGyrA intein to yield Ub-MES (**4**). **d.** Fluorescent SDS-PAGE gels showing the formation of the Ub-CGK(Fluorescein) ligated product (**6**) from one-pot thiolysis and ligation reactions using the inteins indicated. Coomassie-stained SDS-PAGE gels showing MESNa thiolysis of **e.** an SH3 domain, **f.** an SH2 domain, **g.** eGFP, **h.** and the catalytic domain of PARP1 (PARP$_C$). **i.** Comparison of local N-extein residues in model proteins with the native DnaE sequence.

2.5. Expressed Protein Ligation using contiguous DnaE inteins

The most widely used intein-based technology, Expressed Protein Ligation, exploits *cis*-acting inteins to generate recombinant protein α-thioester derivatives.[56] In principle, any split intein can be artificially fused and then utilized as a *cis*-splicing intein in this application (**1** in Figure 2.10a). Ultrafast split inteins are especially attractive in this regard due to their speed and efficiency. To test this notion, we generated artificially fused variants of Npu, Ava, and Mcht with an N-terminal ubiquitin domain and mutations

of the C-terminal (Block G) asparagine and +1 cysteine to alanine. Upon reaction with the exogenous thiol sodium 2-mercaptoethanesulfonate (MESNa), the fused DnaE inteins were rapidly cleaved to generate the ubiquitin α-thioester, **4**, in a few hours (Figure 2.10b). By contrast, MESNa thiolysis of the commonly used MxeGyrA intein was not complete even after one day under identical conditions (Figure 2.10c). Critically, the fused DnaE inteins were sufficiently fast to allow for a one-pot thiolysis and native chemical ligation reaction with an N-terminal cysteine-containing fluorescent peptide, **5**, to give semi-synthetic protein **6** (Figure 2.10d). Furthermore, these inteins could be used to efficiently generate α-thioesters of four other structurally unique proteins domains with different C-terminal amino acid residues (Figure 2.10e-i). These results demonstrate that fused versions of split DnaE inteins are of general utility for protein semi-synthesis.

2.6. The hyper-actived N-terminal splice junction of ultrafast split inteins

The rapid rate of thiolysis observed for the fused DnaE inteins has mechanistic implications as well as practical ones. One possible explanation for their enhanced reactivity over the MxeGyrA intein is that these inteins drive the N-to-S acyl shift reaction more efficiently, generating a larger population of the reactive linear thioester species **2** (Figure 2.10a). This thioester intermediate is generally thought to be transiently populated in protein splicing, and to our knowledge, it has never been directly observed.[102] Surprisingly, when analyzing the ubiquitin-NpuDnaE intein fusion by reverse phase HPLC, we often observed two major peaks and a third minor peak, all bearing the same mass (Figure 2.11). The relative abundance of these species could be modulated by changes in pH (Figure 2.11a,b), and the two major species were almost

equally populated from pH 4-6 (Figure 2.11b). At low pH or after boiling in strong detergents, peak **1** is the dominant species (Figures 2.11a-c), strongly suggesting that peak **1** is the amide, which would be more stable in unfolded protein, and peak **2** is the linear thioester intermediate. Furthermore, we speculate that the minor peak (marked with an asterisk) is the tetrahedral oxythiazolidine intermediate. Importantly, multiple peaks were also visible for the Ub-MchtDnaE and Ub-AvaDnaE fusions (Figure 2.11d,e) but only a single HPLC peak was seen for the Ub-MxeGyrA fusion under identical conditions (Figure 2.11f). These observations, along with the enhanced thiolysis rates, strongly support the notion that these DnaE inteins have a hyper-activated N-terminal splice junction when compared to the MxeGyrA intein.

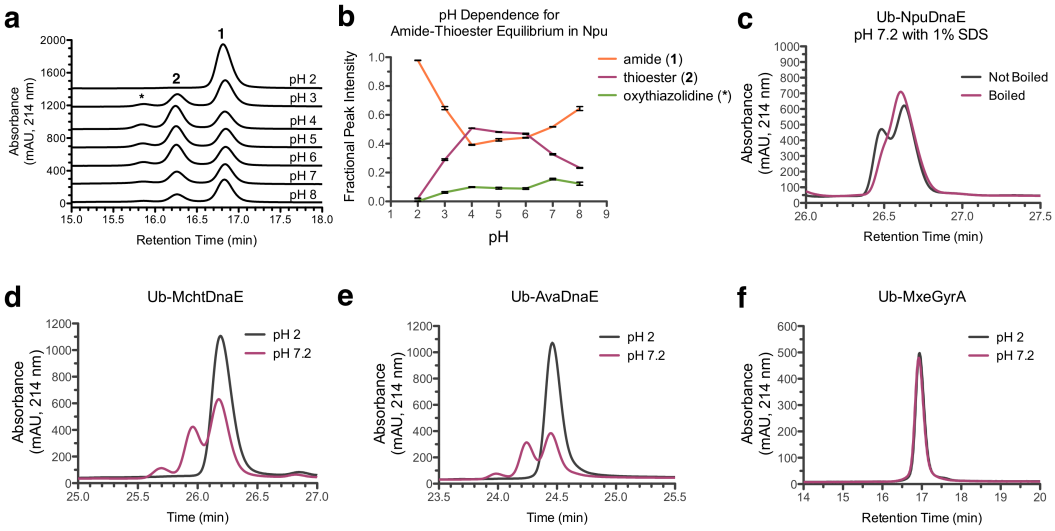


Figure 2.11. Evidence for a highly activated N-terminal splice junction. a. RP-HPLC analysis of Ub-NpuDnaE after pre-incubation in various pH buffers. **b.** Quantification of peak areas from panel a. **c.** RP-HPLC analysis of Ub-NpuDnaE in a pH 7.2 buffer containing 1% SDS, before and after boiling the sample. **d.-f.** RP-HPLC analysis of Ub-fusions to **d.** MchtDnaE, **e.** AvaDnaE, and **f.** MxeGyrA at pH 2 and pH 7.2.

2.7. Streamlined Expressed Protein Ligation (sEPL) using split DnaE inteins

The experiments described above clearly indicate that contiguous variants of the ultrafast DnaE inteins can expedite the EPL procedure by improving the kinetics of

thioester formation. However, isolation of pure protein thioesters from contiguous inteins, including the fused DnaE inteins and MxeGyrA intein, can still be challenging for two reason: (1) All contiguous inteins are active as soon as they are translated and folded, making them susceptible to premature hydrolysis *in vivo* or during initial isolation from cell lysates. In fact, the more activated contiguous DnaE inteins are highly susceptible to this premature cleavage (Figure 2.12). (2) Even after isolation of a protein of interest as an intein fusion, thiolysis may be marred by background hydrolysis of the product thioester. Often, these caveats necessitate the development of customized purification regimes, involving multiple chromatographic steps, to isolate the desired product from complex mixtures.[30,103,104] Collectively, these technical issues mean that a considerable investment in time and resources is usually required before a semi-synthetic protein is obtained in useful quantities.

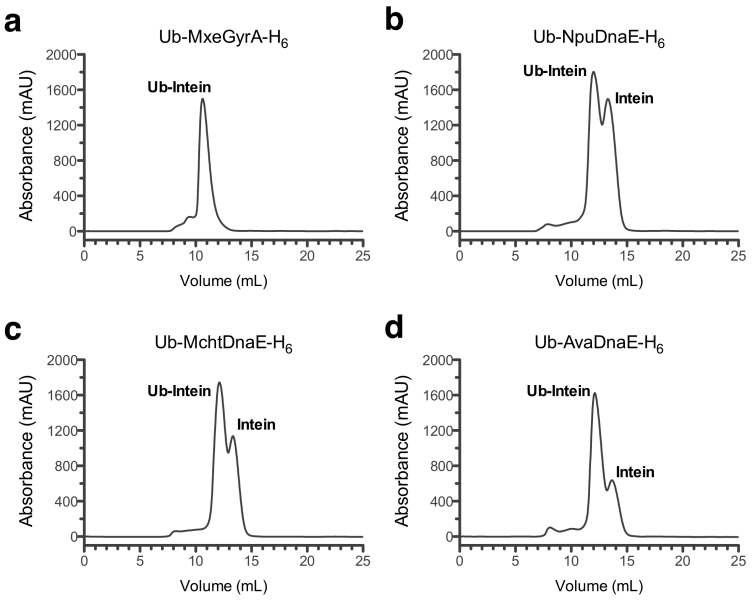


Figure 2.12. Premature cleavage of Ub-intein fusions. His_6-tagged ubiquitin fusions to four contiguous inteins, **a.** MxeGyrA, **b.** NpuDnaE, **c.** MchtDnaE, and **d.** AvaDnaE, all bearing the Block G Asn-to-Ala and +1 nucleophile-to-Ala mutations, were expressed in *E. coli* and enriched over Ni columns. The enriched proteins were injected onto a Superdex 75 column to separate the intact fusion from the prematurely cleaved intein.

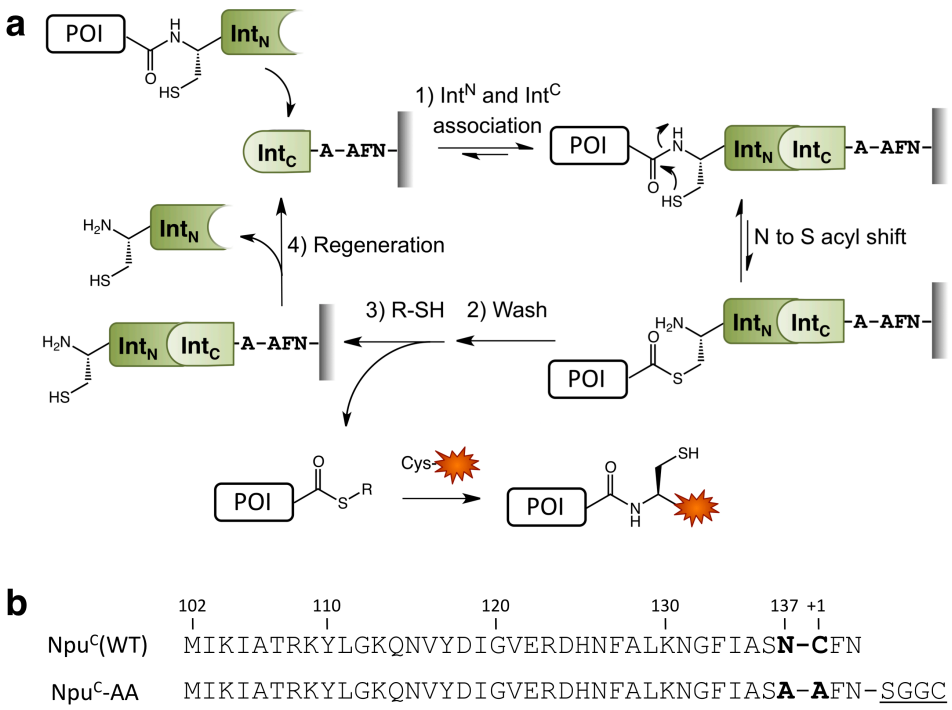


Figure 2.13. Streamlined EPL (sEPL) using split inteins. a. Isolation of an α-thioester derivative of a protein of interest (POI) using engineered split inteins. The product can be directly used for EPL, and the Int$_C$ column can be regenerated. **b.** Sequences of wild-type Npu$_C$ and the Npu$_C$-AA mutant in the Int$_C$ column. Mutated residues are shown in bold, and the linker sequence added for immobilization onto the solid support is underlined.

To overcome the various drawbacks associated with the intein thiolysis process central to EPL, we envisioned an alternative strategy that takes advantage of two important properties of the naturally split form of the DnaE inteins: (1) Cognate N- and C-intein pairs bind tightly to one another; dissociation constants in the low nanomolar range, reflecting extremely fast on-rates, have been reported.[50,98] This ability of split inteins to self-associate was recently exploited by Lu et al. as part of a traceless protein purification system, in this case using an artificially split intein pair.[105] (2) The potential utility of split DnaE inteins in EPL is further enhanced by the fact that the highly activated N-terminal splice junction (Figure 2.11) should only be reactive after the split intein fragments associate.[58,106] Thus, we conceived the integrated protein modification

system shown in Figure 2.13a in which the split intein association is employed both to purify the desired protein from complex biological mixtures and to trigger the generation of a desired protein α-thioester for EPL. In principle, this complementation system should address the major issues attendant to the standard EPL protocol and provide rapid access to highly pure protein α-thioesters.

2.7.1. Split intein-mediated thiolysis and EPL

To implement our system, we designed a variant of the Npu_C peptide in which the C-terminal asparagine and the +1 C-extein residue were mutated to alanine (N137A and C+1A), analogous to our contiguous DnaE constructs. Furthermore, a new cysteine residue was engineered near the C-terminus of the peptide and used to immobilize this fragment on a solid support (Figure 2.13b). The resulting Int_C column was then evaluated as an affinity-modification resin. Three test proteins, maltose binding protein (MBP), ubiquitin (Ub), and protein histidine phosphatase type 1 (PHPT1) were genetically fused to Npu_N and expressed in *E. coli*. In each case, cells were lysed, and the soluble fraction was loaded onto the Int_C column to allow binding of the Npu_N tagged protein to the immobilized Npu_C. After a brief incubation (~5 minutes at room temperature) the column was extensively washed to remove contaminants, and thiolysis was triggered by addition of a buffer containing the thiol MESNa. In all three cases, the desired α-thioester protein eluted from the Int_C column with high recovery (75−95%) and high purity (~95% as determined by RP-HPLC and mass spectrometry) (Figure 2.14a-c). Total isolated yields of purified protein α-thioesters varied from one protein to another and ranged from 2.5 mg (per L of bacterial culture) for Ub-MES to 40 mg for MBP-MES. The calculated

loading capacity of the Int_C column used in these experiments was 3−6 mg of protein per mL of beads (0.12 μmol/mL), but higher or lower loadings could easily be achieved by modifying the amount of Npu_C immobilized on the solid support.

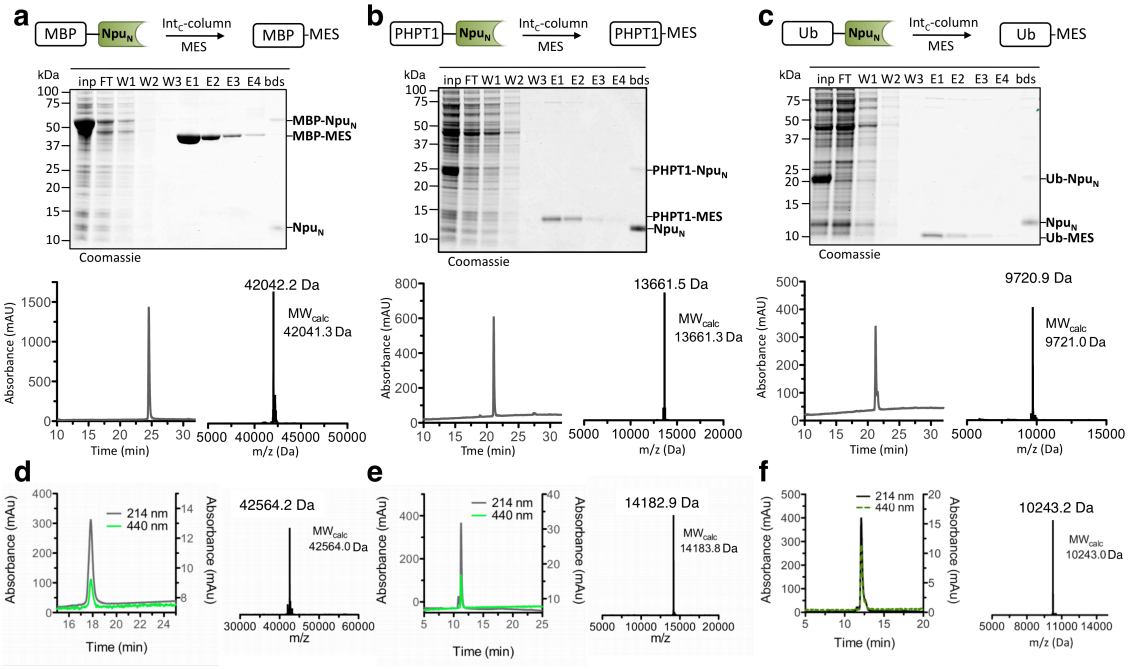


Figure 2.14. Purification of reactive α-thioesters from lysates using split inteins. a. MBP-, **b.** PHPT1-, and **b.** Ub-MES α-thioesters were generated and purified in one step from *E. coli* cell lysates using the Int_C column. The purifications were monitored by SDS-PAGE with coomassie staining (top panels) (inp: input, FT: column flow-through, W1-3: washes, E1-4: elutions, and bds: resin beads). RP-HPLC (detection at 214 nm) and ESI-TOF MS analysis of the eluted fractions (bottom left and right panels, respectively) confirmed the identity of all protein α-thioesters and indicated high purity. The resulting thioesters were reacted with the CGK(Fl) tripeptide to demonstrate EPL compatibility. The products, **d.** MBP-CGK(Fl), **e.** PHPT1-CGK(Fl), and **f.** Ub-CGK(Fl), were analyzed by RP-HPLC (detection at 214 and 440 nm) and ESI-MS to confirm identity and purity. Note that the ligated product Ub-CGK(Fl) (**f**) was isolated by direct incubation of the bound protein on the Int_C column with MESNa and the CGK(Fl) protein to carry out the thiolysis and ligation in one pot.

The utility of the α-thioester derivatives of Ub, MBP, and PHPT1 obtained from the column was demonstrated by ligating each of them to the N-terminal Cys-containing fluorescent peptide (CGK(Fl)) to give the corresponding semi-synthetic products in excellent yield (Figure 2.14d-f). Importantly, one-pot thiolysis/ligation reactions could be

carried out, which allowed us to obtain a site-specifically modified protein directly from cell lysates without isolating the intermediate thioester (Figure 2.14f).

2.7.2. Dependence on the identity of the C-terminal amino acid

An attractive feature of EPL is that it can allow for the preparation of site-specifically modified proteins in a virtually traceless manner. This is contingent on the ability to efficiently generate recombinant protein α-thioesters when any of the 20 proteinogenic amino acids are present at the C-terminus of the protein. The activity of split inteins is known to be sensitive to the identity of the amino acids immediately flanking the splice junction.[90,91] Thus, we were eager to test the generality of our strategy and asked whether we could generate α-thioesters of all 20 amino acids, using ubiquitin as the N-extein template. Twenty Ub-Npu_N fusion proteins were individually expressed in *E. coli* and purified over the Int_C column as before. Thiol-induced cleavage yields from the solid support were determined by SDS-PAGE analysis of the eluted and resin beads fractions, and levels of competing side reactions (mainly hydrolysis) were measured by RP-HPLC and ESI-TOF MS (Figure 2.15). The results clearly show that sEPL is highly compatible with most amino acids at this -1 N-extein position, as was also observed for the contiguous DnaE inteins (Figure 2.10). Furthermore, 80−95% of the eluted material for most amino acids was the desired α-thioester product.

The only significant exceptions to overall recovery were Pro and Glu, for which recovery were 49 and 50%, respectively (Figure 2.15). The Asn mutant displayed high levels of cleavage from the split intein, but almost no α- thioester could be isolated due to side-chain cyclization to form a C-terminal succinimide. A second problematic residue

was Asp, for which we observed some premature cleavage during initial binding to the Int_C resin. Moreover, RP-HPLC analysis of the eluted fractions upon thiolysis revealed two species with the molecular weight of the desired α-thioester. These results were not wholly unexpected, as Asp is known to cleave prematurely from contiguous inteins through side-chain cyclization,[54,107] and its α-thioesters have been reported to migrate to the side-chain carbonyl yielding mixtures of α- and β-isomers.[108] These minor constraints aside, it is clear from these studies that our sEPL system is useful when the majority of amino acids are present as the last residue in the protein of interest.

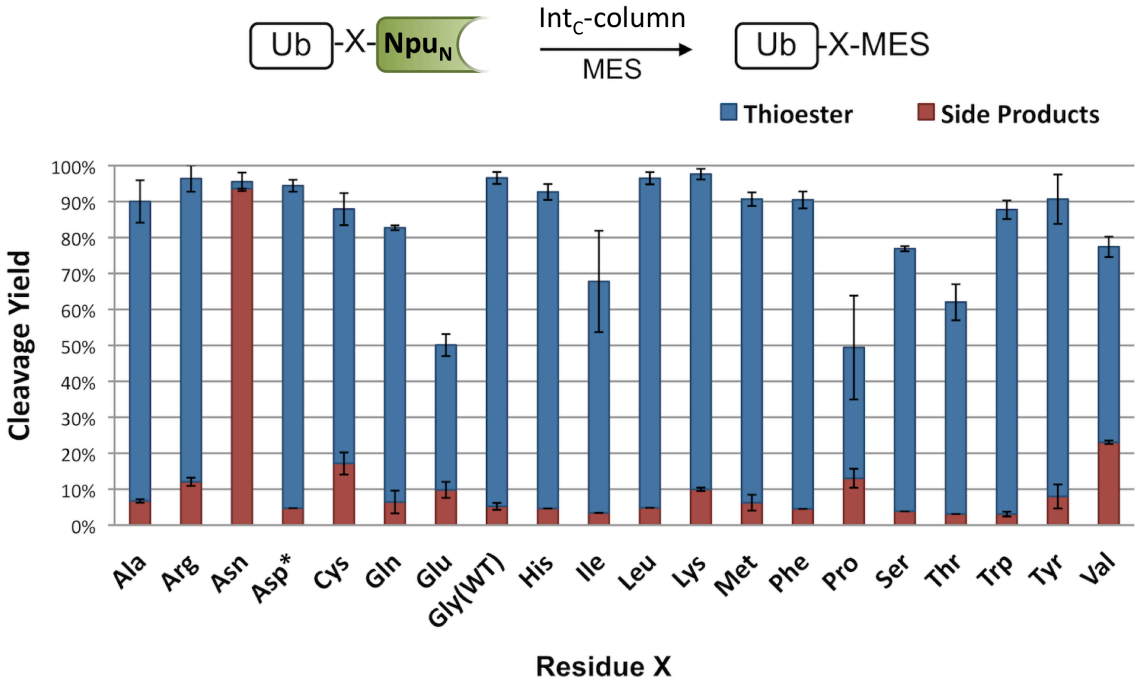


Figure 2.15. Effect of C-terminal amino acid identity on α-thioester formation. The 20 mutants of the protein Ub-X-Npu_N were expressed in *E. coli* varying the identity of the C-terminal amino acid of Ub (X) from the WT Gly to all other proteinogenic amino acids. Thiol-induced cleavage yields from the Int_C column were calculated from the SDS-PAGE analysis of the eluted fractions and left over resin beads. Ratios of α-thioester vs. side products were determined from RP-HPLC and ESI-TOF MS analysis of the eluted fractions. The major competing reaction for all amino acids was hydrolysis with the exception of Asn for which its succinimide form was isolated instead. * See main text for a discussion on the problems associated with Asp. Error bars ± SD (n = 3).

2.7.3. Thioester formation under denaturing conditions

Next we investigated if our system was compatible with denaturing conditions. Protein semi-synthesis frequently requires the preparation of protein fragments, which are often poorly behaved and need to be purified in the presence of strong chaotropic agents. Previous studies have demonstrated that split Npu can splice in the presence of high concentration of denaturants.[89] Thus, we turned to a more challenging target, namely a fragment of histone H2B (residues 1−116), a polypeptide that is prone to aggregation and difficult to generate as an α-thioester derivative using standard EPL procedures.[109] We expressed human histone H2B(1−116) fused to Npu_N in *E. coli*, extracted it from inclusion bodies in 6 M urea, and diluted it to 2 M urea prior to loading on the Int_C-intein column. The incubation was performed for 3 hours at pH 6.0 to maximize binding while avoiding premature cleavage through hydrolysis. The pH was then raised to 7.2, and thiolysis was carried out for 36 hours at room temperature (note that the presence of the denaturant slows down the thiolysis rate). Using these conditions, hH2B(1−116)-MES was obtained in excellent purity (>90% by RP-HPLC) and isolated yield (~20 mg per L of culture) as shown in Figure 2.16a.

This protocol represents a significant improvement over previous approaches which afford less protein (4 mg per L of culture) and require the use of multiple chromatographic purification steps including RP-HPLC.[109] Importantly, the hH2B(1−116)-MES thioester obtained from the Int_C column could be directly used in EPL reactions without further purification. Accordingly, we successfully ligated the protein to a hH2B(117−125) peptide containing an acetylated Lys at position 120[110] to yield semi-synthetic hH2B-K120Ac (Figure 2.16b).

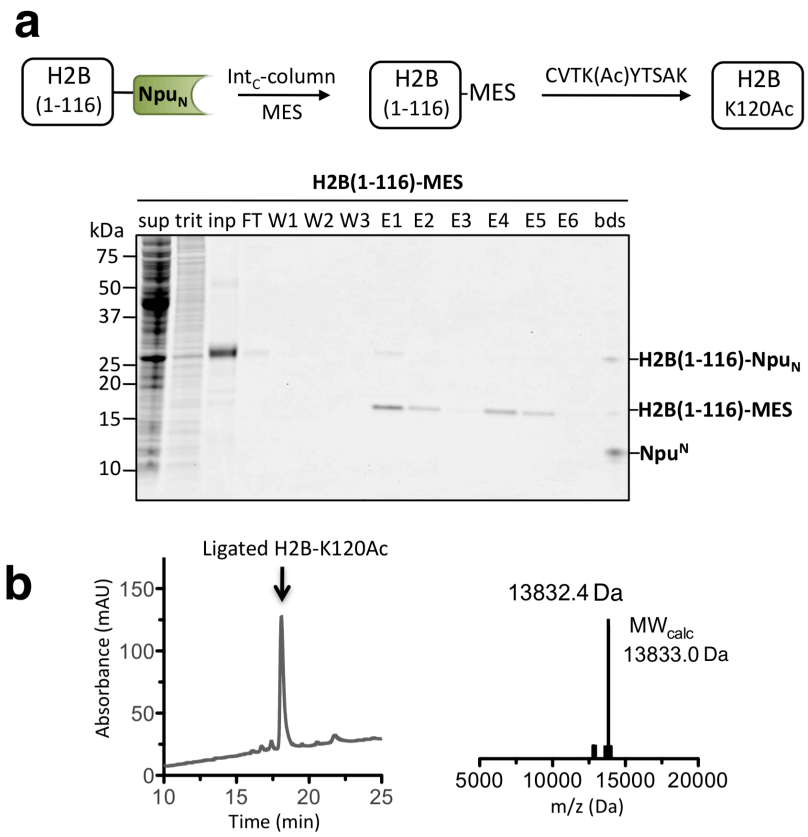


Figure 2.16. Semi-synthesis of hH2B-K120Ac under denaturing conditions. a. Coomassie stained SDS-PAGE analysis of hH2B(1−116) α-thioester generation in the presence of 2 M urea (sup: cell lysate supernatant, trit: 1% triton wash of the inclusion bodies, inp: solubilized inclusion bodies used as input for the Int$_C$ column). E1−E3 were collected after 18 hours of incubation with MES and E4−E6 after an additional 18 hours. E1−E6 were pooled, concentrated to 150 μM, and ligated to the peptide H-CVTK(Ac)YTSAK-OH at 1 mM for 3 hours at room temperature. **b.** RP-HPLC (left) of the ligation reaction mixture and MS (right) of the ligated hH2B-K120Ac product.

2.7.4. Site-specific modification of a monoclonal antibody

Finally, we tested our streamlined EPL methodology for the modification of a monoclonal antibody. The site-specific modification of antibodies has become highly desirable in the area of biopharmaceuticals and diagnostics.[111,112] Currently, most commercially utilized methods to conjugate cargo to antibodies are relatively nonspecific and result in poly-disperse mixtures that may vary from batch-to-batch. Since this heterogeneity can adversely affect both efficacy and safety of the conjugate, attention has

turned toward technologies that afford site-specifically modified antibodies.[113-118] Indeed, protein semi-syntheses via standard EPL and protein *trans*-splicing have recently been used to generate monoclonal antibody conjugates with full activity.[113,119] Given this, we were eager to see whether our streamlined EPL process could be used in the facile generation of antibody conjugates.

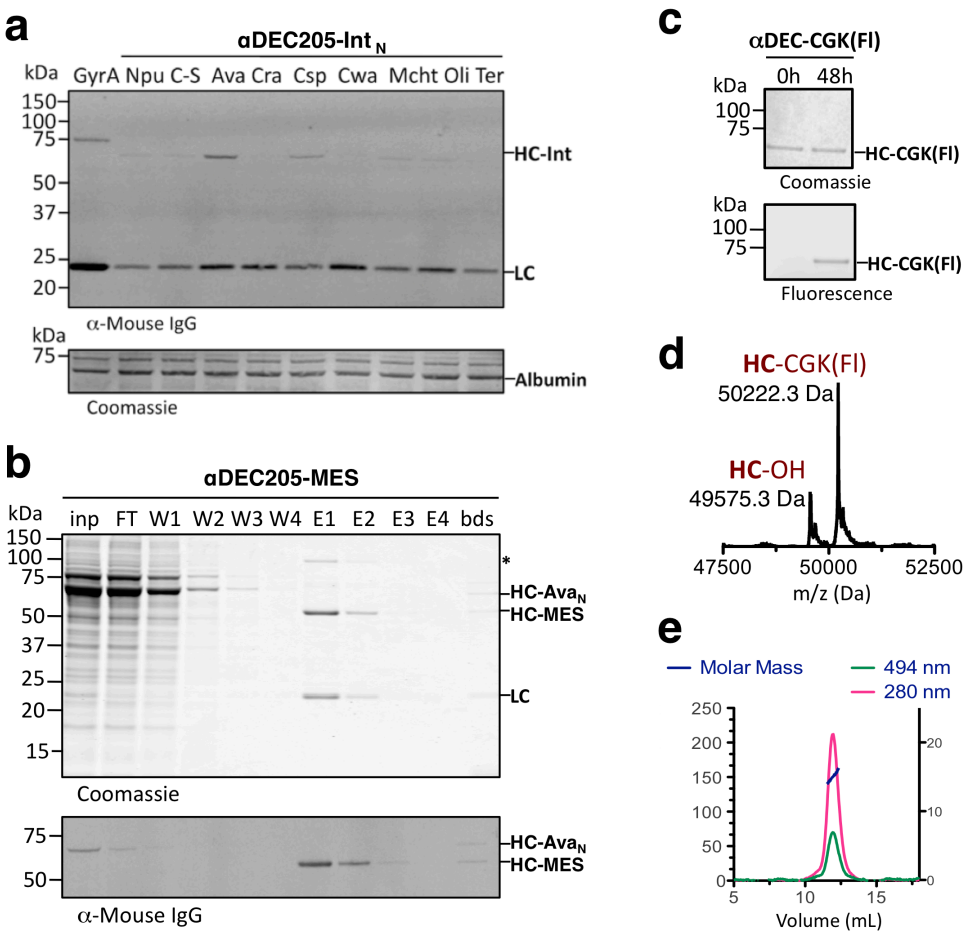


Figure 2.17. Purification and site-specific modification of a monoclonal antibody. a. Western blot of 293T cell supernatants of several αDEC205-Int fusions using an antibody against mouse IgG. A loading control is shown below. **b.** Purification of αDEC205-MES thioester using the split intein column. **c.** Elution fractions containing αDEC205-MES were concentrated to 20 μM and ligated to the CGK(Fl) fluorescent peptide at 1 mM for 48 hours at room temperature. **d.** ESI-TOF MS analysis of degycosylated and fully reduced HC after ligation, showing 75% of the HC is labeled. Expected mass for ligation product = 50221.2 Da. Free HC = 49575.0 Da. **e.** SEC-MALS analysis of the ligated antibody showing that it retains its tetrameric structure after thiolysis and ligation ($MW_{observed}$ = 151 kDa, $MW_{theoretical}$ = 148 kDa).

As a model immunoglobulin (IgG) for our studies, we used an antibody against the DEC205 receptor, a C-type lectin found predominantly on dendritic cells.[120] Accordingly, we designed a construct in which Npu_N was fused to the C- terminus of the heavy chain (HC) of the antibody (αDEC205-Npu_N). Initial expression tests of αDEC205-Npu_N in 293T cells resulted in very low levels of the antibody being secreted (Figure 2.17a). As noted above, the identity of the Int_N can have an effect on expression levels of its fusions (Figure 2.8). Consequently, we asked whether we could obtain higher levels of secreted αDEC205-IntN by varying the identity of the intein N-fragment. Several new αDEC205-Int_N constructs were generated in which Int_N corresponded to the N-fragment of a series of ultrafast split DnaE inteins, namely, Ava, Csp, Cra, Cwa, Mcht, Oli, and Ter. We also tested an Npu_N mutant (C-S) where the noncatalytic cysteines, C28 and C59, were mutated to serine, to determine whether these residues influence secretion and maturation of the IgG tetramer (Figure 2.17a). Importantly, each of the Int_N fragments in this set can cross-react with the C-fragment of Npu without significant loss of splicing efficiency (Figure 2.9). Thus, the Npu_C column already in hand is compatible with all of these αDEC205-Int_N fusions. The contiguous MxeGyrA intein was also fused to αDEC205 to test whether the use of N-intein fragments negatively affected expression levels compared to a full-length intein. An expression screen of this αDEC205-Int_N library was performed in 293T cells, revealing that the Ava_N and Csp_N fusions reproducibly exhibited higher expression levels than the other N-inteins, with the former being the best (Figure 2.17a). Indeed, the expression levels of the αDEC-Ava_N construct were at least as good as the αDEC-GyrA construct.

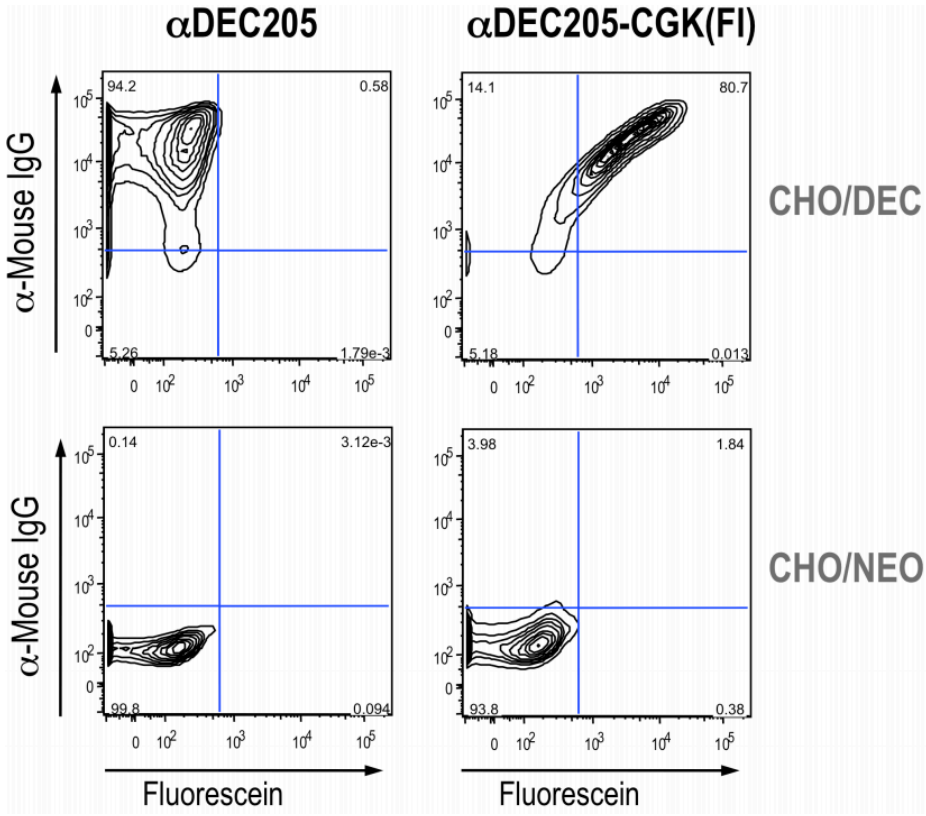


Figure 2.18. Binding of αDEC205-CGK(Fl) to the DEC205 receptor. Contour plots from flow cytometry analysis of the fluorescent, semi-synthetic αDEC205-CGK(Fl) (right panels) or unmodified αDEC205 (left panels) binding to CHO/DEC cells expressing the DEC205 receptor (top panels) or CHO/NEO cells as a negative control (bottom panels). Binding was detected using an α-mouse IgG antibody and through fluorescence of the fluorescein appendage.

Based on this, the αDEC205-Ava_N fusion was chosen and purified over the Int_C column in an analogous manner to that of the soluble proteins described above (Figure 2.17b). Elutions from the column contained the thiolyzed αDEC205, which was subsequently ligated to the CGK(Fl) peptide (Figure 2.17c). MS analysis of the deglycosylated and fully reduced antibody confirmed the formation of a stable, non-reducible amide bond between the αDEC205 heavy chain and the fluorescent peptide with a 75% yield (Figure 2.17d). We also performed size exclusion chromatography (SEC) coupled to multiple angle light scattering (MALS) analysis and confirmed that the

antibody retained its tetrameric folded state after thiolysis and ligation (Figure 2.17e). Importantly, we demonstrated that the semi-synthetic αDEC-CGK(Fl) retains its ability to bind the DEC205 receptor to the same extent as a control αDEC205, previously shown to be fully functional *in vivo* (Figure 2.18).[121] Binding of αDEC-CGK(Fl) to the DEC205 receptor could be monitored by flow cytometry using an anti-mouse IgG secondary antibody and also through the site-specifically incorporated fluorescein (Figure 2.18).

2.8. Summary and conclusions

The results presented in this chapter describe significant progress towards the development of highly optimized protein engineering tools. In the first part, the splicing activities of an entire family of split DnaE inteins were systematically characterized. These experiments showed that ultrafast protein *trans*-splicing is the norm rather than the exception in this family. This represents the first comprehensive comparison of several orthologous inteins, and the data have allowed us to gain the first insights into what molecular interactions are required for rapid protein splicing. Further investigations in this or other intein families to extract more sequence-activity correlations will allow us to better understand the sequence constraints for protein splicing and hopefully lead to the development of even more versatile and useful inteins.

Perhaps the most important discovery presented herein is the observation that not only can several DnaE inteins splice proteins in tens of seconds, but they are particularly enhanced in their capacity to activate a peptide bond for nucleophilic attack in the first step of protein splicing. This fact has clear implications for the improvement of the most prevalent protein semi-synthesis technique, Expressed Protein Ligation. As shown above,

replacing the standard MxeGyrA intein with an artificially fused DnaE intein can greatly enhance the kinetics of protein thioester formation. Furthermore, simply retaining the split form of these inteins can improve EPL to yet another degree, providing a way to purify proteins from complex biological mixtures and generate C-terminal thioesters in a controlled fashion with high purity all at once. This streamlined iteration of EPL is extremely versatile, given that it can be applied to substrates bearing virtually any C-terminal amino acid and to proteins with a wide array of biochemical properties, including histones and monoclonal antibodies.

The work presented in this chapter primarily serves to expand the scope and utility of split inteins for protein engineering. It also provides further support for the idea that split inteins can be used to enhance protein chemistry techniques that traditionally utilize contiguous inteins (as discussed in Chapter 1, section 1.4). Despite this fact, the experiments described herein do not address one known shortcoming of the ultrafast Npu intein: its activity is strongly dependent on local C-extein residues. While this dependence does not effect its utility for EPL or sEPL, which only involve the N-extein, it can hinder the use of Npu for other techniques such as segmental labeling, protein/peptide cyclization, or *in vivo* protein semi-synthesis. In the next chapter, this detrimental property is assessed across the whole DnaE intein family. Then the magnitude and structural origin of C-extein dependence are determined, thus providing a means to potentially engineering C-extein promiscuity into this class of inteins.

Chapter 3: The structural role of the C-extein in protein *trans*-splicing[*]

While different families of inteins utilize subtle variations on the same general biochemical mechanism (such as Ser or Thr nucleophiles, rather than Cys), the catalytic residues for protein splicing are always confined to the intein domain and the first C-extein residue.[22] Despite this fact, a growing body of experimental evidence indicates that intein splicing efficiency is highly dependent on the identity of two or three local extein residues on either side of the splice junction.[54,90-92,122] For example, introduction of non-native residues at the -3, -2, and -1 positions, located on the N-extein (Figure 3.1), can alter the linear thioester formation efficiency or promote hydrolysis of this intermediate. Mutation of the +1, +2, and +3 residues, located on the C-extein (Figure 3.1), can abolish or greatly diminish splicing activity and even lead to premature asparagine cyclization before branched intermediate formation. For each intein family, this context dependent activity is dictated by evolutionary pressures, as inteins are naturally embedded between highly conserved residues in a number of different endogenous host proteins.[11] As a result, different inteins are biased towards different sequences at their splice junction.

[*] The experiments described in section 3.1 were carried out with Geoffrey Dann. The remainder of the chapter describes the results of a close collaboration with Ertan Eryilmaz in Professor David Cowburn's lab at Albert Einstein College of Medicine. The work in this chapter was published in the following two papers:

Shah, N. H., Dann, G. P., Vila-Perelló, M., Liu, Z., and Muir, T. W. Ultrafast protein splicing is common among cyanobacterial split inteins: Implications for protein engineering. *J. Am. Chem. Soc.* **134**, 11338-11341 (2012).

Shah, N. H., Eryilmaz, E., Cowburn, D., and Muir, T. W., Extein residues play an intimate role in the rate-limiting step of protein trans-splicing. *J. Am. Chem. Soc.* **135**, 5839-5847 (2013).

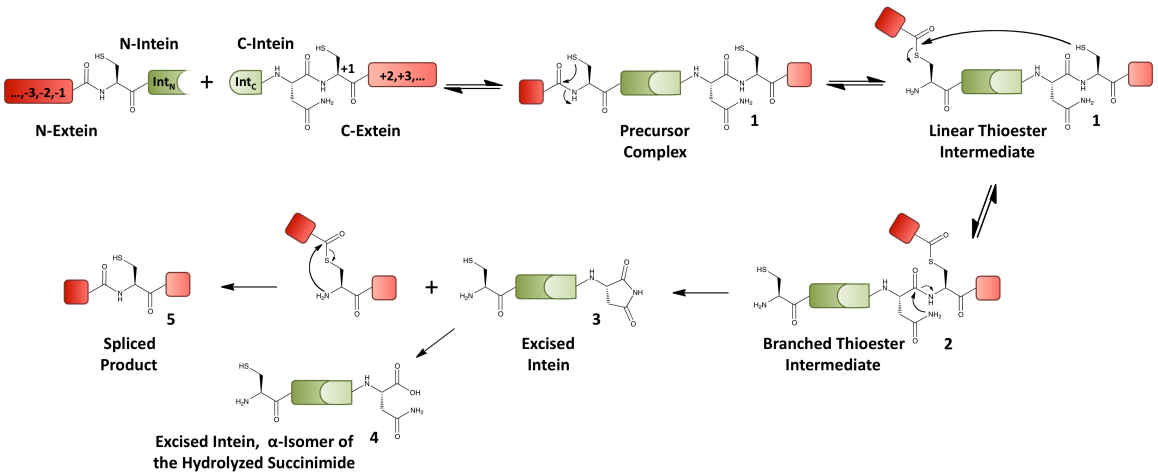


Figure 3.1. The mechanism of protein *trans*-splicing (PTS). Relevant species along the reaction coordinate are labeled. Numbers **1**-**4** refer to the chemically distinct C-intein adducts that can be observed in the *in vitro* splicing assays described in this chapter (see Figure 3.4). Note that for simplicity, only the α-amino acid isomer of **4** is shown, however species **3** can ring-open into an α- or β-amino acid form.

The chemical synthesis of larger and more complex peptides and proteins is an ongoing challenge, and inteins are being widely used to facilitate such syntheses.[8] Thus, the sensitivity of protein splicing to local extein sequence (i.e. residues immediately flanking the intein) has significant practical implications. All intein-based technologies are premised on a single notion: the chemical perturbations that an intein carries out on its endogenous host protein can be applied in a virtually traceless manner to any exogenous protein of interest. In reality, however, efficient and traceless synthesis of complex products is not always achieved. Rather, current technologies often require either the incorporation of non-native residues surrounding the splice junction in the target molecule or the sacrifice of reaction kinetics and product yields to obtain the desired native sequence. An improved understanding of the general splicing mechanism and of its sensitivity to local extein sequences thus remains of central concern.

Of particular interest as protein engineering tools are the split DnaE inteins, all of which endogenously generate the catalytic subunit of DNA polymerase III after protein

trans-splicing.[19] Until recently, many split intein-based technologies relied on the founding member of this family termed Ssp, which derives its name from the model cyanobacterium that encodes it, *Synechocystis species PCC6803*. However, Ssp catalyzes protein *trans*-splicing in hours, which is too slow for many practical applications.[88] With the discovery and characterization of new split DnaE inteins, such as the now prevalent *Nostoc punctiforme* (Npu) intein, it is clear that several members of this family catalyze protein splicing with extraordinary efficiency, in minutes or less.[89,90,123] Thus, many intein technologies are now being developed and improved with these new tools, including *in vitro* and *in vivo* protein semi-synthesis,[86,101,124] segmental isotopic labeling,[93,100] peptide cyclization,[125] and the construction of novel biosensors.[126,127]

All split DnaE inteins are naturally embedded within the local N-extein sequence AEY (Figure 3.1, residues -3, -2, -1) and C-extein sequence CFN (Figure 3.1, residues +1, +2, +3). Several reports indicate that DnaE inteins can tolerate significant deviation from this native N-extein sequence.[89,91,123,124,128] Conversely, the presence of non-native C-extein residues can lead to dramatic reductions of splicing efficiency in both the Npu and Ssp inteins. For example, mutation of the canonical CFN sequence to SGV inhibits branched intermediate resolution for Ssp, although the contributions of each C-extein mutation were not individually assessed.[92] Additionally, the identity of the +2 C-extein residue has a dramatic impact on splicing activity for both of these inteins, but it is not clear what step in the splicing pathway is modulated by this residue.[90]

Despite the fact that C-extein-dependent splicing activity is well documented for Npu and Ssp, little is known about how conserved this effect is, the magnitude of this effect on reaction kinetics, or the physical basis of this phenomenon. We envisioned that

a detailed understanding of how C-extein residues participate in the splicing reaction could help guide the practical use of split inteins and help lay the foundation for the design of more promiscuous engineered inteins. To this end, we first surveyed C-extein dependence in the entire DnaE family to assess conservation of this phenomenon. Next, we performed a detailed structure–activity analysis on the Npu intein, employing semi-synthesis to systematically alter the C-extein moiety, thereby providing the raw materials for a series of kinetic and structural analyses. This effort led to the finding that the +2 residue in the C-extein plays a critical role in constraining the active site of the intein during resolution of the branched intermediate. The work also draws attention to a loop region in the intein structure that appears to sense C-extein composition and as such might be a productive focus of engineering efforts geared towards increasing intein promiscuity (discussed further in Chapter 6).

3.1. Preliminary survey of C-extein dependence in DnaE inteins

Since the activities of Npu and Ssp are known to be dependent on the identity of the +2 C-extein residue (natively phenylalanine), we first sought to determine if other related inteins share this property. Thus, we analyzed all of the split DnaE inteins in the presence of a +2 glycine (CGN), glutamic acid (CEN), or arginine (CRN) in the *in vivo* screening assay described in Chapter 2 (Figure 2.2). Importantly, this assay can be carried out in the background of varying local C-extein sequences without significantly perturbing the dynamic range (Figure 3.2a).[92] Like Npu and Ssp, most of the inteins showed a dramatic decrease in activity in the presence of all three +2 mutations (Figure 3.2b). Of the tested amino acids, glutamic acid was tolerated best for every intein, suggesting a conserved mechanism for accommodating a negative charge at this position.

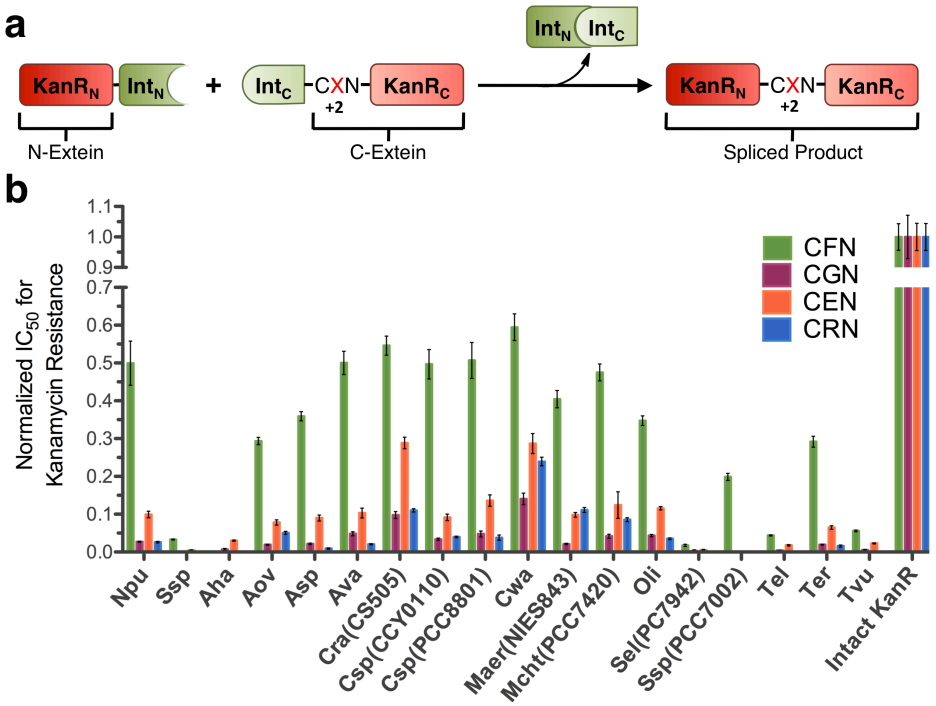


Figure 3.2. *Trans*-splicing of split DnaE inteins *in vivo* with +2 mutations. a. Scheme depicting intein activity-dependent kanamycin resistance. **b.** Relative *trans*-splicing efficiencies at 30°C with the endogenous "CFN" C-extein sequence and exogenous "CGN", "CEN", and "CRN" sequences were determined using the assay depicted in Figure 2.2. IC_{50} values (± SE, n = 3-4) are normalized to the intact KanR proteins with the corresponding tri-peptide.

In an attempt to quantitatively assess the magnitude of the effect of C-extein mutations on *trans*-splicing, we analyzed the Npu, Cra(CS505), and Cwa inteins using our *in vitro* SDS-PAGE kinetic assay (Figure 2.4a) in the presence of a +2 glycine. All three of these reactions were characterized by rapid accumulation of the branched thioester intermediate, which slowly resolved over tens of minutes into the spliced product and the N-extein cleavage product (Figure 3.3). Consistent with previously reported observations, these data indicate that split DnaE inteins require steric bulk at the +2 position for branched intermediate resolution and efficient splicing.[92] While the effect of the F+2G mutation was clear in all three cases, the instability of the branched intermediate during gel sample preparation and the prevalence of side reactions in this

assay made quantitative kinetic analysis challenging. Thus, we developed a modified kinetic assay that utilized semi-synthetic inteins bearing short exteins where *trans*-spliicng could be monitored by reverse phase high performance liquid chromatography (RP-HPLC) or electrospray ionization mass spectrometry (ESI-MS).

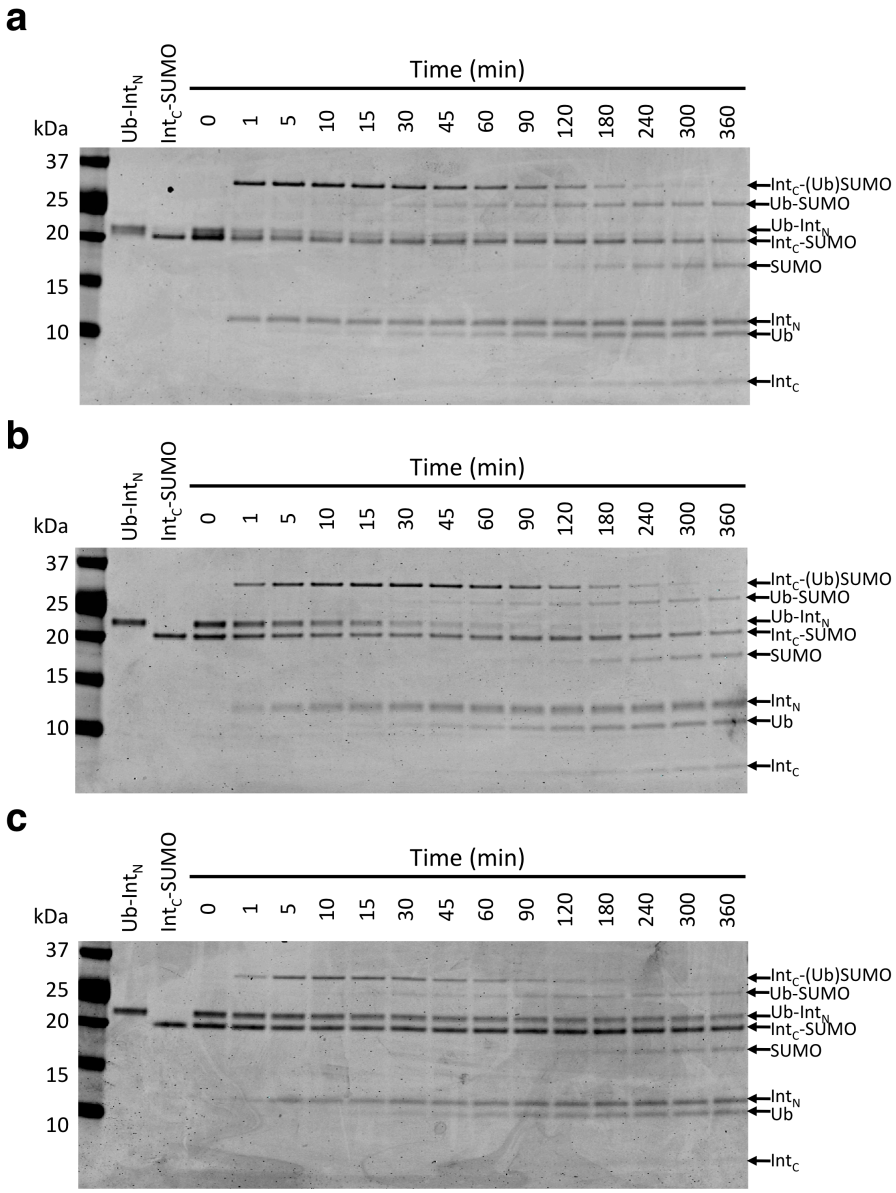


Figure 3.3. *In vitro* SDS-PAGE splicing assays with a +2 glycine residue. SDS-PAGE analysis of **a.** Npu, **b.** Cra(CS505), and **c.** Cwa *trans*-splicing reactions. Gel bands were visualized by coomassie staining. Note that in order to observe the branched intermediate, Int_C-(Ub)SUMO, gel samples were prepared in SDS loading dye lacking thiols, acidified before boiling, then neutralized prior to loading the gel.

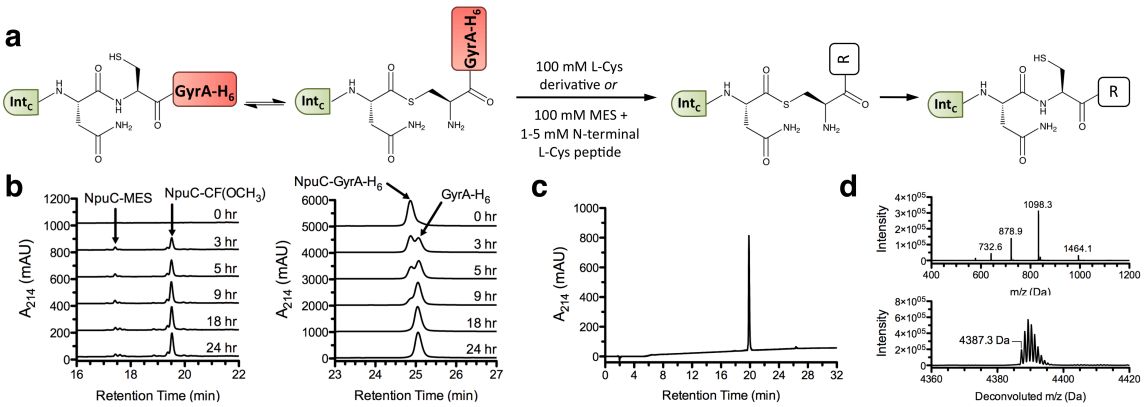


Figure 3.4. Semi-synthesis of C-intein constructs. a. Semi-synthetic scheme (R = -OH, $-OCH_3$, $-NH_2$, $-NHCH_3$, or an additional one or two amino acids, as indicated in Table 3.1). **b.** RP-HPLC analysis of a one-pot MES-thiolysis/ligation to synthesize Npu_C-$CF(OCH_3)$. Npu_C-MES and ligation product accumulation are shown in the left panel and cleavage of the Npu_C-GyrA-H_6 fusion protein is shown in the right panel. **c.** RP-HPLC and **d.** ESI-MS analysis of Npu_C-$CF(OCH_3)$ after purification. The raw mass spectrum is shown in the top panel and the deconvoluted spectrum is shown in the bottom panel (expected monoisotopic mass = 4387.28 Da).

3.2. Semi-synthesis of split inteins with varying C-extein composition.

Our efforts began with the construction of a library of C-intein fragments (Int_C) bearing a variety of model C-exteins ranging from a single Cys residue with different capping groups to tri-peptides with unique sequences. To rapidly generate the desired constructs, seventeen proteins in all, we employed a semi-synthetic approach that utilized Expressed Protein Ligation (Figure 3.4).[33] Specifically, the Int_C fragments of Npu and Ssp (referred to as Npu_C and Ssp_C, respectively) were expressed in *E. coli* fused to the *cis*-splicing His_6-tagged GyrA intein and enriched over Ni columns. The crude fusion proteins were then reacted with either a large excess of a cysteine derivative (100 mM) to directly yield an Int_C-Cys adduct, or they were thiolyzed with 100 mM 2-mercaptoethanesulfonate (MES) in the presence of a 1-5 mM di- or tri-peptide to yield Int_C-peptide adducts (Figures 3.4a,b). The desired product from each reaction was readily purified by reverse-phase high performance liquid chromatography (RP-HPLC, Figures

3.4c) and its identity was confirmed by electrospray ionization mass spectrometry (ESI-MS, Figure 3.4d). Importantly, this semi-synthesis approach allowed for the modular assembly of constructs with natural amino acid mutations within the C-intein and effectively any functional groups in the C-extein side-chains and backbone.

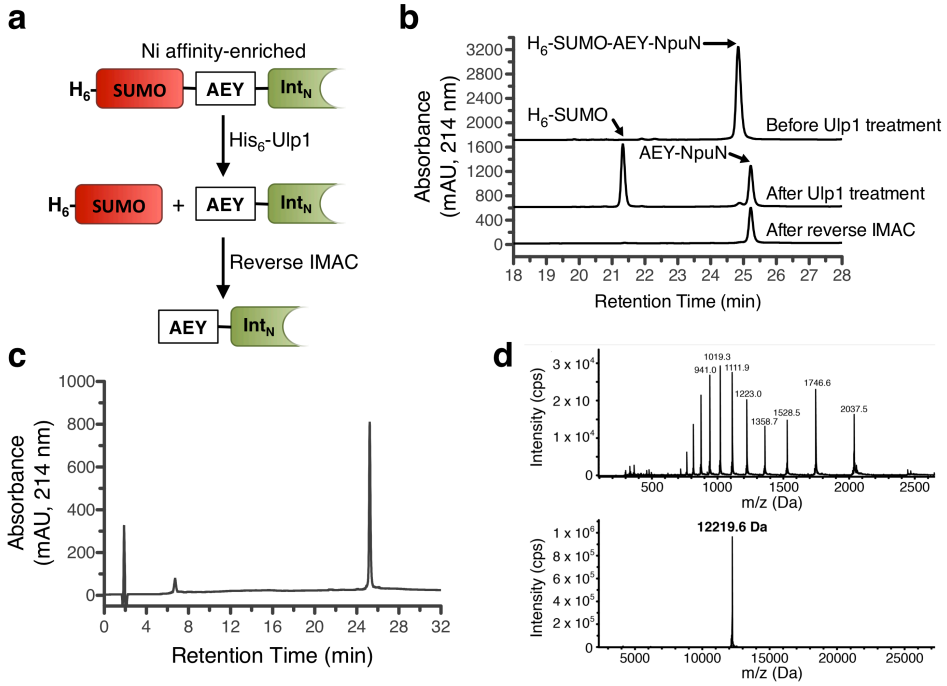


Figure 3.5. Purification of NpuN with a minimized N-extein. a. Scheme for processing and isolation of a recombinant N-intein with the "AEY" N-extein. **b.** RP-HPLC analysis of SUMO cleavage and immobilized metal affinity chromatography (IMAC) depletion of undesired materials. **c.** RP-HPLC and **d.** ESI-MS after further purification by size-exclusion chromatography. The raw mass spectrum is shown on top and the deconvoluted spectrum is shown on the bottom (expected average mass = 12219.79 Da).

3.3. Kinetic assays to monitor branched intermediate formation and resolution.

To rigorously assess C-extein effects on protein *trans*-splicing, we developed two complementary analytical approaches that allowed us to distinguish various chemical species along the reaction coordinate in a time-resolved fashion. First, N-intein (Int_N) proteins bearing a minimized N-extein tripeptide (AEY-Int_N) were generated recombinantly and purified (Figure 3.5). These constructs were mixed with their Int_C

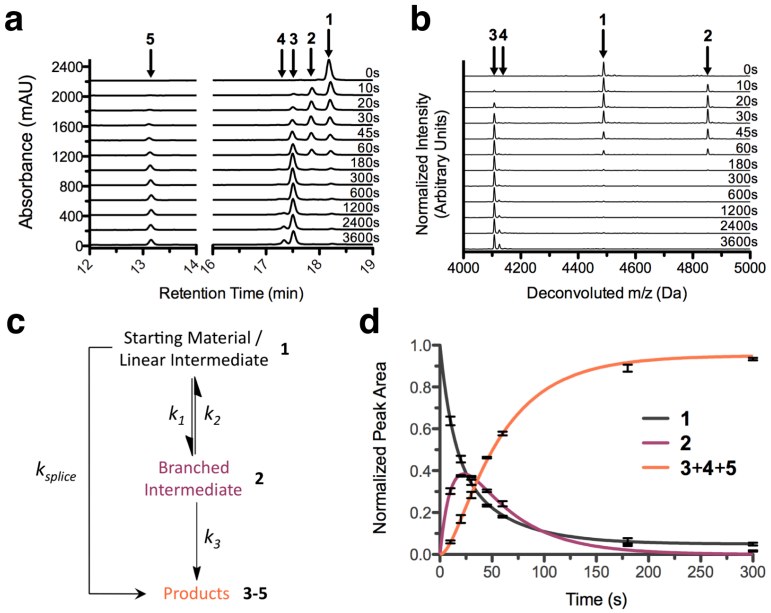


Figure 3.6. Splicing assays to analyze branched intermediate formation and resolution. Time-dependent **a.** RP-HPLC and **b.** ESI-MS analyses of the reaction between AEY-Npu_N and Npu_C-CFN(NH_2) (Table 3.1, reaction 1). The deconvoluted mass spectra in panel b are normalized to the intensity of the largest peak at each time point. **c.** Simplified three-state kinetic model of protein splicing compatible with the analytical techniques presented herein. **d.** Quantified reaction progress data for reaction 1 in Table 3.1 fit to the kinetic model in panel c. Note that the product curve is the combined peak areas of all three products of the *trans*-splicing process, the two excised intein species (**3** and **4**) and the spliced product (**5**). Error bars represent the standard deviation from three independent reactions. Numbering corresponds to the numbered species defined in Figures 3.1.

counterparts at 30 °C, and aliquots were removed from the reaction solution at various time points and quenched by acidification to pH 1-2. Importantly, all reactions were carried out at pH 7.2 in the absence of thiol-based reducing agents to prevent any undesired hydrolysis or thiolysis reactions that would convolute kinetic analyses. Time points were analyzed by RP-HPLC, and for most reactions the various Int_C-related species (Figure 3.1, **1**-**4**) and the spliced product (Figure 3.1, **5**) could be readily separated (Figures 3.6a). For reactions where sufficient separation between species **1**-**5** was not achieved by RP-HPLC, the quenched time points were desalted and analyzed as

complex mixtures by ESI-MS (Figures 3.6b). Given the similarity in sequence composition, size, and net charge between species **1**-**4**, the molecules showed similar levels of ionization, and thus the RP-HPLC analyses and ESI-MS analyses gave virtually identical results (compare Figures 3.6a and 3.6b). Importantly, in both assay formats, the starting material and linear intermediate were indistinguishable (since the C-intein construct is chemically unchanged in these two states, Figure 3.1), so the data were fit to a simplified kinetic model that collapsed the first two catalytic steps into a single equilibrium reaction (Figures 3.6c,d). The results of our kinetic analyses are summarized in Table 3.1 and Figure 3.7.

We initially carried out a series of control reactions to validate our assays. The splicing kinetics of the wild-type Npu and Ssp inteins were assessed in their native N- and C-extein contexts (Table 3.1, reactions 1 and 2). The overall rates of spliced product formation (k_{splice}) were 1.36 x 10^{-2} s^{-1} and 1.46 x 10^{-4} s^{-1}, respectively, consistent with previous measurements from gel-based assays.[88,89,123] These experiments also demonstrated that the slowest step, BI resolution (described by k_3), is rate-determining for Ssp but the initial and latter steps of PTS are kinetically coupled for the faster Npu reaction. As additional controls, we independently mutated the first catalytic cysteine, Cys_1, and the C-terminal asparagine, Asn_{137}, in Npu to alanine and analyzed the effect of these mutations on splicing activity. As expected, the C1A mutation completely inhibited splicing, however a basal level of succinimide formation, and thus C-extein cleavage, was observed on a time scale of hours (Table 3.1, reaction 3). This result is consistent with the notion that C-terminal asparagine cyclization is stimulated by branched intermediate formation, as was previously shown for the GyrA intein.[30] Additionally, the

N137A mutation abolished splicing and C-extein cleavage, but only modestly reduced the kinetics of the initial steps (Table 3.1, reaction 4).

Table 3.1. Rate constants for individual steps and the overall splicing reaction.[a]

Rxn	Intein	C-Extein	k_1 (s^{-1})	k_2 (s^{-1})	k_3 (s^{-1})	k_{splice} (s^{-1})
1	Npu_{WT}	$CFN(NH_2)$	$(5.21 \pm 0.28) \times 10^{-2}$	$(1.77 \pm 0.38) \times 10^{-2}$	$(3.15 \pm 0.04) \times 10^{-2}$	$(1.36 \pm 0.02) \times 10^{-2}$
2	Ssp_{WT}	$CFN(NH_2)$	$(4.70 \pm 0.26) \times 10^{-3}$	$(7.03 \pm 0.44) \times 10^{-3}$	$(3.86 \pm 0.17) \times 10^{-4}$	$(1.46 \pm 0.03) \times 10^{-4}$
3 [b]	Npu_{C1A}	$CFN(NH_2)$	-	-	$(1.43 \pm 0.03) \times 10^{-4}$	-
4 [c]	Npu_{N137A}	$CFN(NH_2)$	$(1.70 \pm 0.13) \times 10^{-2}$	$(1.86 \pm 0.11) \times 10^{-3}$	-	-
5 [d]	Npu_{WT}	C(OH)	$(2.41 \pm 0.07) \times 10^{-2}$	$(4.40 \pm 0.14) \times 10^{-3}$	-	-
6	Npu_{WT}	$C(OCH_3)$	$(6.59 \pm 0.20) \times 10^{-2}$	$(1.56 \pm 0.06) \times 10^{-2}$	$(4.76 \pm 0.13) \times 10^{-4}$	$(4.32 \pm 0.16) \times 10^{-4}$
7	Npu_{WT}	$C(NH_2)$	$(3.16 \pm 0.09) \times 10^{-2}$	$(6.59 \pm 2.41) \times 10^{-3}$	$(7.31 \pm 0.26) \times 10^{-5}$	$(6.30 \pm 0.87) \times 10^{-5}$
8	Npu_{WT}	$C(NHCH_3)$	$(4.40 \pm 0.35) \times 10^{-2}$	$(1.13 \pm 0.10) \times 10^{-2}$	$(1.33 \pm 0.01) \times 10^{-4}$	$(1.08 \pm 0.03) \times 10^{-4}$
9	Npu_{WT}	$CF(OCH_3)$	$(5.90 \pm 0.85) \times 10^{-2}$	$(1.20 \pm 0.39) \times 10^{-2}$	$(1.56 \pm 0.16) \times 10^{-3}$	$(1.28 \pm 0.01) \times 10^{-3}$
10	Npu_{WT}	$CF(NH_2)$	$(6.10 \pm 1.10) \times 10^{-2}$	$(1.13 \pm 0.40) \times 10^{-2}$	$(9.30 \pm 0.42) \times 10^{-3}$	$(6.32 \pm 0.11) \times 10^{-3}$
11	Npu_{WT}	$CFA(NH_2)$	$(6.05 \pm 0.36) \times 10^{-2}$	$(1.31 \pm 0.27) \times 10^{-2}$	$(2.57 \pm 0.04) \times 10^{-2}$	$(1.31 \pm 0.02) \times 10^{-2}$
12	Npu_{WT}	$CAN(NH_2)$	$(7.11 \pm 2.11) \times 10^{-2}$	$(2.74 \pm 0.58) \times 10^{-2}$	$(3.12 \pm 0.28) \times 10^{-4}$	$(2.39 \pm 0.12) \times 10^{-4}$
13 [b]	Npu_{C1A}	$CAN(NH_2)$	-	-	$(2.41 \pm 0.02) \times 10^{-6}$	-
14	Npu_{H125N}	$CFN(NH_2)$	$(4.21 \pm 0.46) \times 10^{-2}$	$(8.96 \pm 3.22) \times 10^{-3}$	$(5.53 \pm 0.50) \times 10^{-4}$	$(4.92 \pm 0.13) \times 10^{-4}$
15 [e]	Npu_{H125N}	$CAN(NH_2)$	$(7.81 \pm 0.34) \times 10^{-2}$	$(2.94 \pm 0.01) \times 10^{-2}$	$(3.23 \pm 0.27) \times 10^{-5}$	$(3.23 \pm 0.27) \times 10^{-5}$
16	Npu_{D124Y}	$CFN(NH_2)$	$(7.75 \pm 0.55) \times 10^{-2}$	$(2.06 \pm 0.23) \times 10^{-2}$	$(3.27 \pm 0.11) \times 10^{-2}$	$(1.74 \pm 0.07) \times 10^{-2}$
17	Npu_{D124Y}	$CAN(NH_2)$	$(1.06 \pm 0.76) \times 10^{-1}$	$(3.87 \pm 0.47) \times 10^{-2}$	$(4.43 \pm 0.05) \times 10^{-4}$	$(3.61 \pm 0.21) \times 10^{-4}$

[a] k_1, k_2, and k_3 were extracted from a global fit of all three normalized curves for one reaction to the analytical solutions for the differential rate equations that describe our kinetic model. k_{splice} was extracted by fitting the product formation curve to a standard first-order rate equation. The values represent the average and standard deviation from three individually fit unique reactions.

[b] In reactions 3 and 13, the mutation of Cys_1 precludes the first steps of the splicing pathway. k_3 represents the rate of succinimide formation and thus C-extein cleavage in the absence of branch formation.

[c] In reaction 4, mutation of the catalytic asparagine abolishes succinimide formation, thus the reaction does not progress past the branched intermediate.

[d] In reaction 5, while all catalytic residues are present, no branched intermediate resolution was observed during the course of the assay.

[e] The extremely slow BI resolution in reaction 15 led to roughly 10-20% N-extein hydrolysis as a side reaction, preventing global fitting to our kinetic model. For this reaction, k_1 and k_2 were extracted from a two-state equilibrium kinetic model using only the pre-equilibrium phase of the reaction (first 10 minutes). k_3 was assumed to be identical to k_{splice}, which was determined by fitting the product formation curve to a first order rate equation.

3.4. C-extein effects on branched intermediate formation and resolution.

Next, we employed our kinetic assays to determine the effect of C-extein composition on individual steps in the *trans*-splicing reaction (Table 3.1, reactions 5-12). These experiments revealed that C-extein variation had only a small effect on the kinetics of BI formation (k_1 and k_2), while it profoundly affected the BI resolution step (k_3) and thus the overall splicing rate (k_{splice}) (Figures 3.7a,b). A detailed comparison of these kinetic analyses revealed several important trends (Figure 3.7c). First, C-extein chain-

length had a substantial effect on activity. Cys_{+1} alone could not sustain BI resolution with an uncapped carboxylate, suggesting that a negative charge near the active site is undesirable (Table 3.1, reaction 5). Capping the +1 residue as an amide or an ester restored a basal level of splicing activity (Table 3.1, reactions 6-8). Interestingly, Cys_{+1} capped with a methyl ester afforded a 4-fold rate increase over the methyl amide analog, possibly indicating an inhibitory role for this amide N-H moiety or an anomalous non-native effect of this subtle perturbation (Table 3.1, reactions 6 and 8). Ultimately, the effect of chain-length on BI resolution was more pronounced once the entire Phe_{+2} residue was added (Table 3.1, reaction 10), however three C-extein residues were required to recapitulate the fastest reported rates for Npu (Table 3.1, reaction 1).

Through our kinetic analyses we also identified two specific functional groups that made major contributions to BI resolution. First, we found that the amide bond after Phe_{+2} provided a 6-fold rate enhancement relative to a methyl ester (compare reactions 9 and 10). This result suggests that the amide N-H group is involved in a hydrogen bond that facilitates BI resolution, for example, by stabilizing a catalytically competent conformation. The second, more significant functional group is the Phe_{+2} phenyl ring. While this residue is known to be important, as discussed above, the extent of its contribution to BI resolution was not previously known. Our measurements indicate that the addition of the bulky Phe side-chain enhances BI resolution kinetics 100-fold relative to Ala (compare reactions 1 and 12). Interestingly, the presence of the Phe side-chain also stimulated the basal rate of succiminide formation (i.e. C-extein cleavage) 60-fold relative to Ala in the context of the C1A mutation (Table 3.1, compare reactions 3 and 13), implying that the Phe side-chain is making favorable interactions even in the absence

of the BI. By contrast, the side-chain of Asn_{+3} does not contribute to the *trans*-splicing reaction (compare reactions 1 and 11).

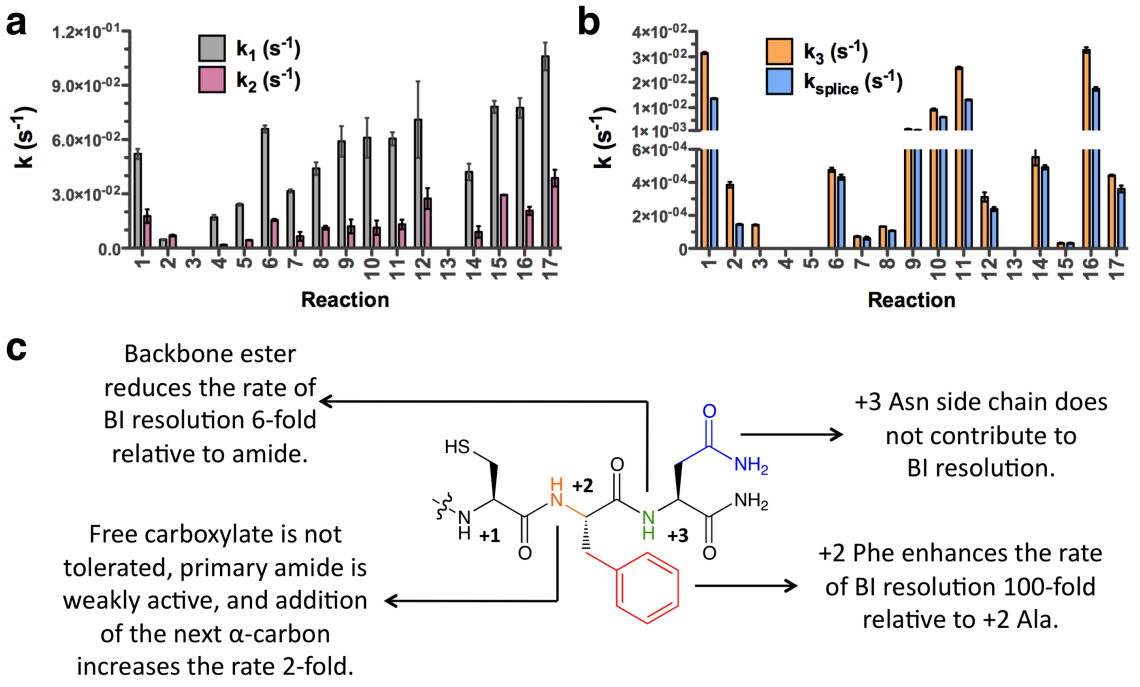


Figure 3.7. C-Extein contributions to splicing activity. a. Forward (k_1) and reverse (k_2) rates of branched intermediate formation from starting materials. **b.** Rate of branched intermediate resolution (k_3) and overall rate of *trans*-splicing (k_{splice}). **c.** Scheme highlighting the key conclusions from the kinetic data.

3.5. A structural role for the +2 C-extein residue.

Given the significant contribution of the Phe_{+2} side-chain to splicing kinetics, we next sought to understand the structural origin of its involvement in split intein chemistry. Most high resolution structures of inteins, including the only two published structures of Npu,[39,49] do not contain native C-extein residues. One important exception to this is a crystal structure of Ssp bearing five native N-extein residues (KFAEY), three native C-extein residues (CFN), and mutations of the terminal intein residues, Cys and Asn, to Ala.[38] In this structure, the Phe_{+2} side-chain packs against a catalytic histidine that lies on a flexible loop (Figure 3.8a). This histidine (His_{125} in Npu) is completely conserved in the

DnaE family and has been implicated as a general acid or base in the BI resolution step of many inteins.[30,38] Mutation of His_{125} in Npu to an Asn reduced the rate of BI resolution roughly 60-fold, similar to the F+2A mutation (Table 3.1, reactions 14 and 12, respectively). The Ssp structure suggests that Phe_{+2} participates in protein *trans*-splicing by stabilizing His_{125} through a direct interaction. Indeed, the effect of mutating both residues in Npu on BI resolution kinetics was non-additive ($\Delta\Delta G_{coupling}$ = 1.07 kcal mol^{-1}) indicating some cooperativity between Phe_{+2} and His_{125} with respect to this step (Table 3.1, reaction 15, and Figure 3.8b for thermodynamic cycle analysis).

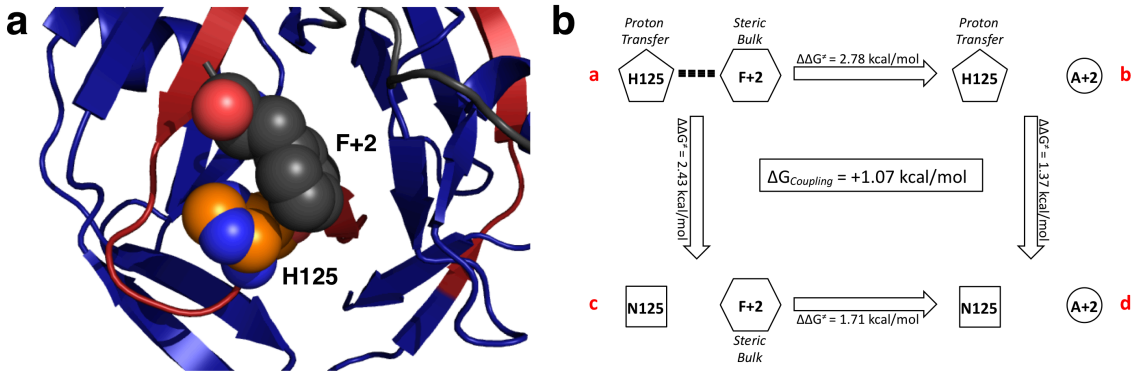


Figure 3.8. Interaction between His_{125} and Phe_{+2}. a. Crystal structure of the SspDnaE intein (pdb: 1ZDE) highlighting close packing of His_{125} and Phe_{+2} in spheres. The N-intein and C-intein are shown as blue and red ribbons, respectively. **b.** Mutant cycle analysis of the His_{125}-Phe_{+2} interaction's contribution to splicing. The coupling free energy indicates a loss of favorable interaction energy for BI resolution upon mutation of His_{125} and Phe_{+2}.

To better understand the structural impact of the +2 residue, we carried out solution NMR analyses of Npu in both a CFN(NH_2) or CAN(NH_2) C-extein context. NMR constructs were prepared analogously to those used for kinetic assays with some additional provisions. Specifically, the Npu_N protein contained the native N-extein sequence (AEY) and an inactivating C1A mutation, but was not ^{13}C or ^{15}N isotopically labeled. The Npu_C constructs, bearing the N137A mutation, were ^{13}C and ^{15}N enriched in

the recombinant Int_C portion, but not in the synthetic C-extein region. N- and C-inteins were mixed, and the complexes were purified to homogeneity by size exclusion chromatography. Use of this segmental labeling scheme meant that only the Npu_C residues (Ile_{103}-Ala_{137}), which have identical chemical composition in both complexes, would be visible in heteronuclear correlation experiments. This was expected to simplify assignment whilst still allowing the putative interaction between the +2 residue and the catalytic His_{125} to be interrogated. The inactivating mutations (C1A and N137A) ensured that chemistry would not occur during data acquisition.

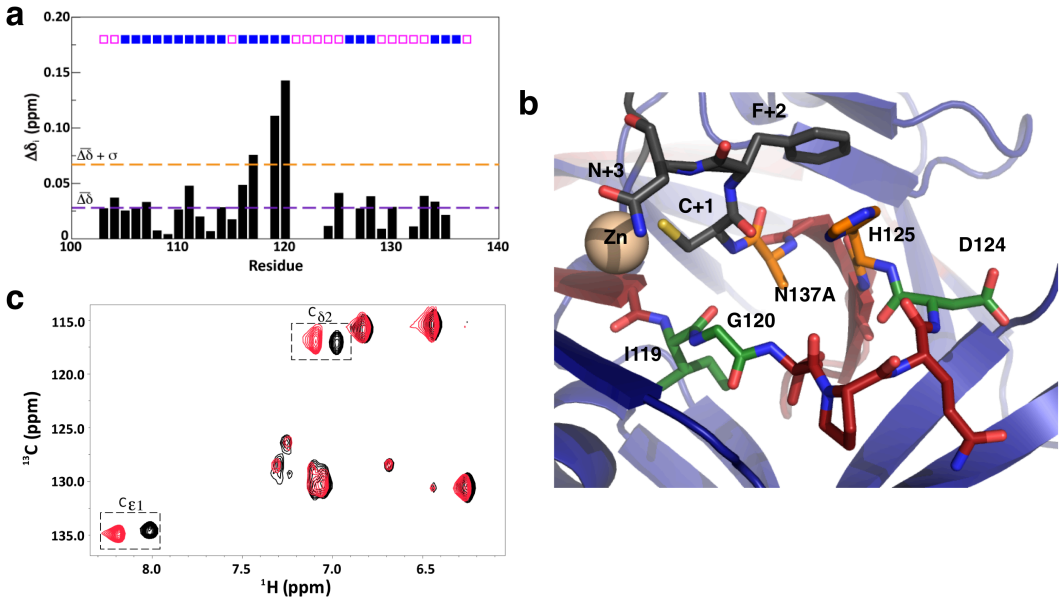


Figure 3.9. Structural effects of mutating the C-extein +2 residue. a. Composite ^{1}H and ^{15}N backbone chemical shift perturbations ($\Delta\delta_i$) in Npu_C (^{13}C,^{15}N labeled) in complex with unlabeled Npu_N as a function changing the +2 C-extein residue from Phe to Ala (see methods for calculations). The mean value is marked by a dashed purple line, and one standard deviation above the mean is marked by a dashed orange line. Residues in secondary structure elements are marked with boxes above the bars, solid blue boxes are strands and empty pink boxes are loops. **b.** The active site of Ssp bearing catalytic Cys and Asn mutations, native extein residues, and a coordinated zinc ion. Important residues surrounding the C-intein/C-extein junction (orange/black junction) are shown as sticks. The C-extein is shown in gray, key catalytic residues are shown in orange, and other non-catalytic residues highlighted in this study are shown in green. **c.** Overlay of the aromatic region of the ^{1}H-^{13}C HSQC spectra of the Npu_N : Npu_C complexes containing either Phe (black) or Ala (red) as the +2 C-extein residue. Chemical shift perturbations of His_{125} imidazole ring ^{1}H-^{13}C correlations, $C\varepsilon_1$ and $C\delta_2$, are marked in dashed boxes.

With the exception of three residues (121-123) in the loop containing the catalytic His_{125} residue, we were able to assign the majority of the Npu_C backbone resonances in the complexes using standard triple-resonance experiments. Most of the backbone resonances were unperturbed upon changing the +2 C-extein residue from Phe to Ala (Figure 3.9a). The only exceptions to this were the amide resonances from Ile_{119} and Gly_{120}, which showed a modest perturbation. These residues are located at the beginning of the loop containing the catalytic His residue and, in the Ssp crystal structure, lie close to the peptide bond between N_{137} and C_{+1} that is ultimately attacked during branched intermediate resolution (Figure 3.9b). The His_{125} backbone amide resonance was not itself sensitive to the nature of the +2 C-extein residue. However, the aromatic side-chain protons of this residue did exhibit significant chemical shift perturbations on mutating the +2 residue, suggesting an altered chemical environment for this side-chain in the absence of the +2 phenyl ring (Figure 3.9c). Together with our mutagenesis and kinetic data, these NMR studies lend support to the idea that the active site conformation of Npu is coupled to the identity of the C-extein +2 residue.

3.6. The Phe_{+2} C-extein residue constrains active site motions.

In order to gain additional insight into the interplay between C-extein residues and the Npu active site, we carried out molecular dynamics (MD) simulations of two wild-type intein complexes bearing either $CFN(NH_2)$ or $CAN(NH_2)$ as C-exteins (identical to the constructs in Table 3.1, reactions 1 and 12). Simulations were carried out in explicit solvent in 1 fs steps for 0.5 μs. Comparison of the two simulation trajectories afforded a more detailed picture of the coupling between the +2 residue and the intein active site

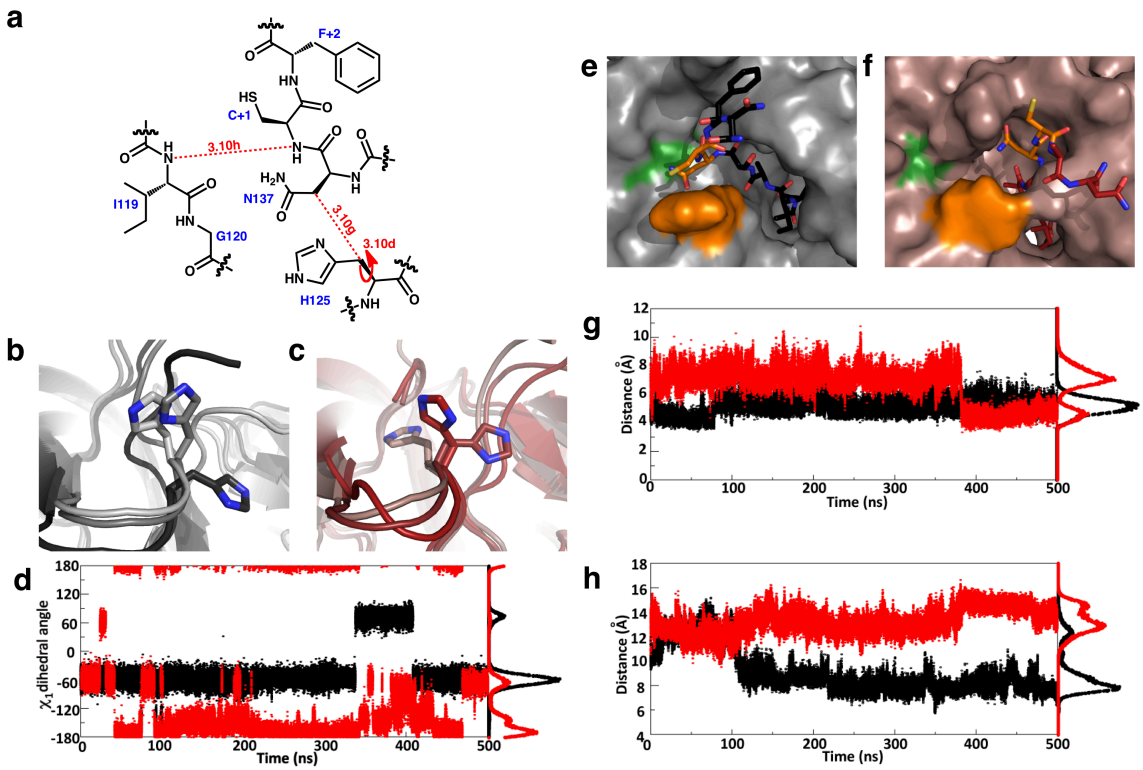


Figure 3.10. Molecular dynamics simulations to probe +2 amino acid-dependent active site dynamics. a. Scheme depicting measurements extracted from MD simulations. The distance or angular measurement is shown in read, labeled by the panel where the corresponding trajectory is shown. Three representative frames from the MD trajectories of **b.** AEY-Npu_N + Npu_C-CFN(NH_2) and **c.** AEY-Npu_N + Npu_C-CAN(NH_2) highlighting His_{125} rotameric states. **c.** Trajectory of the His_{125} side-chain dihedral angle (χ_1) during the simulations. Representative frames highlighting the positioning of Asn_{137} relative to the His_{125} loop in **e.** the CFN simulation and **f.** the CAN simulation. Asn_{137} and Cys_{+1} are shown as orange sticks, His_{125} is shown as an orange surface, and Ile_{119} and Gly_{120} are shown as green surfaces. **g.** Distance between His_{125} and Asn_{137} Cβ atoms during the simulations. **h.** Distance between Ile_{119} and Cys_{+1} amide nitrogens during the simulations. Data from the simulation with the CFN(NH_2) C-extein are shown in black, and analogous data from the simulation with the CAN(NH_2) C-extein are shown in red. Traces to the right of trajectory graphs are histograms indicating the distribution of angles or distances sampled throughout the simulation.

(Figure 3.10). One of the more striking results from the simulation was the effect of changing +2 C-extein on the dynamics of the His_{125} side-chain. In the presence of Phe_{+2} the His side-chain primarily adopts a single rotameric state with only a brief excursion to an alternate rotamer (Figures 3.10b,d, black trajectory). By contrast, with an Ala_{+2}

residue, His_{125} frequently switches between three side-chain rotamers and favors a different conformation than the one found with Phe_{+2} (Figures 3.10c,d, red trajectory). The backbone ϕ and φ dihedral angles for His_{125} showed no significant change as a function of C-extein composition. These data are consistent with the fact that there were chemical shift perturbations for the His_{125} side-chain but not the backbone.

The second major consequence of the +2 residue mutation was the overall positioning of the C-intein/C-extein junction (i.e Asn_{137}-Cys_{+1}) relative to the His_{125} loop. In the simulation with CFN as the C-extein, Asn_{137} remained buried in the groove above this loop, similar to the Ssp structure (Figure 3.10e). By contrast, in the CAN simulation, the entire strand bearing Asn_{137} and the C-extein occupied space outside of this groove region (Figures 3.10f). An important consequence of this is that the distance between Asn_{137} and His_{125} (Figure 3.10g) and between Ile_{119} and the scissile peptide bond (Figure 3.10h) were significantly shorter for the majority of the CFN simulation than for the CAN simulation. Overall, these MD simulations indicate that the presence of a sterically bulky amino acid at the +2 position in the C-extein acts to constrain the motions of key catalytic residues leading to a more compacted arrangement around the scissile peptide bond.

In considering the mechanistic implications of these observations it is important to emphasize that, by necessity, the simulations employed a linear precursor protein as the starting point. Use of a BI structure in the simulations would have been more desirable given that our kinetic data reveal that formation of this intermediate stimulates cleavage of the peptide bond at the C-intein/C-extein junction (Table 3.1, compare reactions 1 and 3). Unfortunately, there is currently no high-resolution structural

information on any split DnaE intein in the branched intermediate state. Thus, we were forced to extrapolate from the structures available. Despite this caveat, the major conclusion from the simulation work is broadly consistent with our mutagenesis and kinetic data. In particular, we observe coupling between the +2 residue and the catalytic His_{125} both in the simulations and in the kinetics of BI resolution. We further note that the Phe side-chain stimulates C-extein cleavage even in the absence of the BI (Table 3.1, compare reactions 3 and 13), arguing that this bulky side-chain augments catalysis even in the linear precursor.

3.7. An activating point mutation on the His_{125} loop.

Local C-extein residues appear to affect the structure and dynamics of residues surrounding the flexible His_{125} loop, thereby modulating BI resolution kinetics. Thus, it is conceivable that point mutations within the intein that alter loop conformation or flexibility could also modulate splicing activity and even tolerance to non-native extein residues. In a previous directed evolution study on an Npu_N-Ssp_C chimera, we identified several mutations that make this intein more tolerant of the C-extein sequence SGV, rather than CFN.[92] Intriguingly, one of these mutations was an Asp-to-Tyr mutation adjacent to His_{125} (Asp_{124}). We found that this mutation enhanced the rate of Npu splicing by 50% in the presence of Ala_{+2} (Figure 3.11a and Table 3.1, compare reactions 12 and 17). Importantly, this mutation was still tolerated when Phe_{+2} was present, suggesting that it increases overall promiscuity towards C-exteins (Table 3.1, compare reactions 1 and 16). The Npu NMR structure[39] and the Ssp crystal structures[38,48] indicate that Asp_{124} packs against a β-turn from the N-intein. Given this close packing, the bulky D124Y

mutation would require conformational rearrangement and possibly also rigidification of the catalytic His_{125} loop, which can modulate activity. As predicted, in a 100 ns MD simulation of Npu_{D124Y} with a CAN(NH_2) C-extein, the His_{125} loop conformation was altered, His_{125} rotamer dynamics were constrained, and Asn_{137} persistently remained above the His_{125} loop, similar to the Npu_{WT}-CFN(NH_2) simulation (Figures 3.11b-d). This simulation suggests that the D124Y mutation reduces C-extein dependence by recapitulating the constraints on active site dynamics typically applied by Phe_{+2}, specifically the stabilization of His_{125} and the appropriate positioning of the C-intein/C-extein junction close to His_{125}.

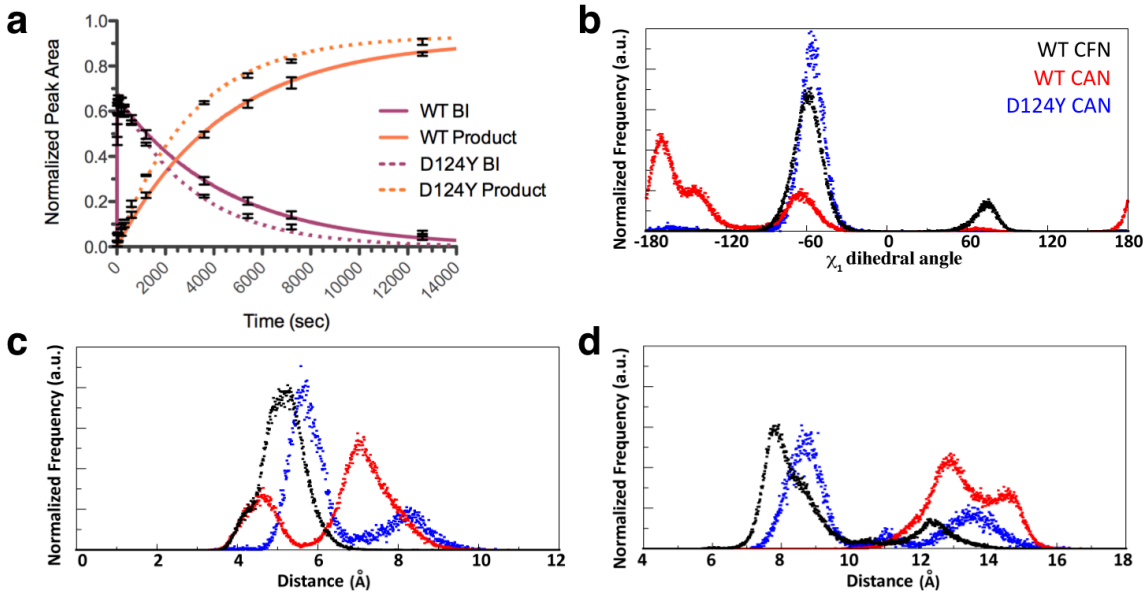


Figure 3.11. Structural and functional effect of the D124Y mutation in Npu. a. Kinetic data showing rate enhancement for the BI resolution step with the D124Y mutation in the CAN(NH_2) C-extein context. Solid lines correspond to best fit kinetic curves for the wild-type intein (Table 3.1, reaction 12), and dotted lines correspond to the D124Y mutant (Table 3.1, reaction 17). Only the BI and product reaction curves are shown, and the starting material curve is omitted for clarity. Error bars represent the standard deviation from three independent reactions. Histograms showing the distribution of **b.** His_{125} χ_1 dihedral angles, **c.** distances between His_{125} and Asn_{137} Cβ atoms, and **d.** distances between Ile_{119} and Cys_{+1} amide nitrogens during the wild-type CFN (black), wild-type CAN (red) and D124Y CAN (blue) simulations.

3.8. Summary and conclusions

In this work, we examined the molecular determinants for C-extein-dependent protein *trans*-splicing. This investigation was facilitated by the utilization of protein semi-synthesis to generate inteins linked to a variety of C-exteins and by the development of novel kinetic assays that provide information about individual steps along the *trans*-splicing reaction coordinate. Through these studies, we not only extracted information on C-extein requirements, but also gained additional mechanistic insights into split DnaE intein splicing. Specifically, our experiments confirmed that branched intermediate resolution is the slowest step for PTS (k_3). They also provided evidence supporting the notion that some DnaE inteins have a highly activated N-terminal splice junction ($k_1/k_2 >$ 2 for all Npu constructs), consistent with our previous report (see section 2.6).[123] Interestingly, this N-terminal activation appears roughly ten-fold slower and is significantly less efficient ($k_1/k_2 = 0.67$) for the Ssp intein. Additionally, we found that for Npu, the rate of Asn cyclization upon BI formation is 200-fold faster than its rate in the absence of the branched structure. Stimulation of Asn cyclization upon BI formation is also found in the *cis*-splicing GyrA intein.[30] We propose that this kinetic stimulation is a common feature of inteins, in effect creating a trigger that helps ensure the proper fidelity of the reaction by minimizing premature cleavage of the C-extein. Lastly, it is particularly surprising that the H125N mutation does not completely abolish BI resolution, but rather reduces its rate 60-fold. Indeed, the splicing rate of this mutant is still faster than for wild-type Ssp. For many non-DnaE inteins, this step requires two histidine residues, one analogous to His_{125} and another immediately preceding the C-terminal Asn residue.[29,30] Given the lack of this penultimate histidine in the DnaE inteins,

His_{125} has been implicated as the sole general acid/base for BI resolution.[38] Our data suggest that while His_{125} is clearly important for BI resolution, other unidentified residues must also contribute to catalysis of this step.

The current study improves our understanding of the relationship between C-extein composition and *trans*-splicing efficiency. Our survey of C-extein tolerance in all DnaE inteins indicates that strong dependence on the identity of the +2 C-extein residue is a conserved feature of this family. The kinetic data indicate that the C-extein almost exclusively affects the BI resolution step. Within the C-extein, we identified specific functional groups that contribute significantly to splicing kinetics, in particular the Phe_{+2} side-chain. Our NMR experiments and MD simulations illustrate that this bulky functional group constrains active site motions, forcing catalytic histidine and asparagine residues and the scissile peptide bond in close proximity. The need for a bulky side-chain at the +2 position is further highlighted by a recent genetic selection study on the Npu intein showing that Trp is also well tolerated at this position.[128] Collectively, these data paint a picture of the split DnaE intein active site that effectively extends beyond the intein domain itself to include the +2 C-extein residue.

During protein *trans*-splicing, the N-extein is transferred from the N-terminus of the intein onto a C-extein side-chain, thereby creating a unique branched protein structure. As BI resolution is the slowest and often rate limiting step for many inteins, this structure is most relevant to overall activity. To date, all published high-resolution structural data on inteins examine either a precursor or product form of the intein. While these studies, have provided substantial insights into the structural basis for protein splicing, they cannot examine interactions that are exclusively present in the branched

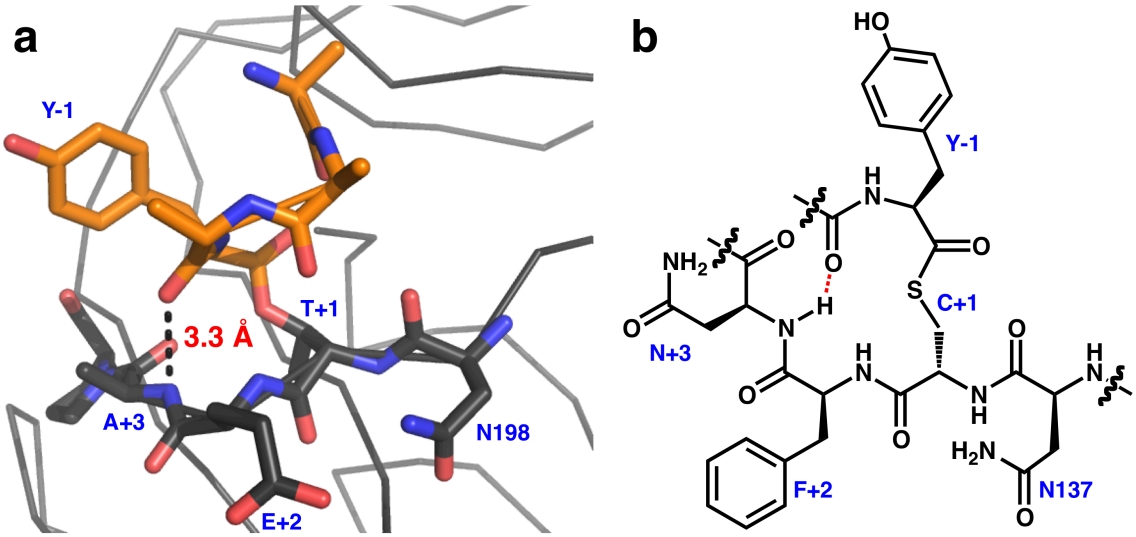


Figure 3.12. Putative hydrogen bond between exteins in the branched intermediate. **a.** Observed hydrogen bond in the MxeGyrA branched intermediate crystal structure. **b.** Hypothesized analogous hydrogen bond in the NpuDnaE branched intermediate.

intermediate. Indeed, our kinetic analyses revealed several important functional groups in the C-extein that affect BI resolution (Figure 3.7c), however only in the case of the Phe_{+2} side-chain could we postulate any kind of structural basis of this from published structures. Recently, our lab solved a high resolution structure of the MxeGyrA intein trapped in its branched intermediate state.* While the native local N- and C-extein sequences for this intein differ from the DnaE family (the N-extein ends with MRY rather than AEY, and the C-extein begins with TEA rather than CFN), this structure shows potentially important backbone interactions involving the exteins that could be conserved in other inteins. For example, as described in section 3.4, our kinetic data indicate that the amide N-H group between the +2 and +3 residues is important for BI resolution (Figure 3.7c and Table 3.1 reactions 9 and 10). In the MxeGyrA BI structure, the analogous

* This structure was determined by Silvia Frutos of the Muir lab in collaboration with Matt Bick from Professor Seth Darst's lab at The Rockefeller University. This work is currently unpublished.

amide N-H moiety is involved a hydrogen bond with a backbone amide within the N-extein (Figure 3.12a). This interaction could help organize the active site for BI resolution and, if also present in the Npu BI (Figure 3.12b), could explain the 6-fold decrease in activity that results from converting the amide to an ester (Figure 3.7c). Ultimately, our results reinforce the need for high-resolution structural information on the branched intermediate in DnaE inteins.

The fullest deployment of split inteins in protein engineering ultimately requires a truly traceless *trans*-splicing system with no sequence requirements. While bulky hydrophobic residues other than phenylalanine are tolerated at the critical +2 position for DnaE inteins, thus alleviating some sequence constraints,[90,128] these inteins are still only modestly promiscuous. Our results suggest that the interplay between the C-extein and the His_{125} active site loop has direct implications for the rational design of improved, more extein-tolerant split inteins. Indeed, the D124Y point mutation on this flexible loop increases the tolerance of Npu for a +2 alanine residue without affecting its activity in a native context. In a recent directed evolution endeavor on a DnaB family intein, a mutation at this position was also found to reduce C-extein sequence constraints.[95] Furthermore, we previously demonstrated that mutating other residues on this loop can generally enhance the activity of Ssp (section 2.3)[123] and the Npu_N-Ssp_C chimera[92] in a native C-extein context. These results collectively indicate that the conformational preferences of this loop are intimately linked with inadequate BI resolution both for intrinsically slow inteins and for efficient inteins in an exogenous C-extein context. Thus, this loop is a hot-spot on the intein structure that should be explicitly targeted in future engineering efforts for the design of more high-activity, broad-specificity inteins.

Chapter 4: The role of charge segregation in complex stability and splicing*

Coulombic forces play an important role in facilitating protein-protein interactions. It has been demonstrated that a strong electrostatic potential between two interacting proteins correlates with a fast rate of association and a strong binding affinity.[129] Furthermore, this principle has been exploited to enhance the binding properties of engineered protein interfaces[130] and to modulate specificity in protein-protein interactions.[131-133]

Electrostatic forces have previously been postulated to play a role in the assembly of the split DnaE intein from *Synechocystis* sp. PCC6803 (Ssp), as the fragments if this intein are oppositely charged proteins, and their association rate is ionic strength-dependent.[50] In fact, sequence alignments of all split DnaE inteins show highly conserved charge segregation with acidic residues concentrated to specific positions on the N-intein and basic residues conserved on the C-intein (Figure 4.1a).[96] Interestingly, when the conserved charged residues are mapped onto the DnaE intein structures, many are found to be participating in intermolecular ion pairs and ion triads (Figure 4.1b).[39] Furthermore, a bioinformatic sequence analysis of the intein family indicates that this charge segregation is significantly more prevalent in naturally split inteins than intact ones (Figure 4.1c).

Despite these observations, the importance of specific ionic interactions in split intein assembly has not been experimentally validated. In this study, using site-directed

* The work in this chapter was carried out in close collaboration with Miquel Vila-Perelló and was published in the following paper:

Shah, N. H., Vila-Perelló, M., and Muir, T. W., Kinetic control of one-pot trans-splicing reactions by using a wild-type and designed split intein. *Angew. Chem. Int. Ed.* **50**, 6511-6515 (2011).

mutagenesis coupled to splicing activity assays and equilibrium binding measurements, we experimentally confirm that specific electrostatic interactions are important for split intein fragment association and thus activity. Furthermore, we demonstrate that the degree of charge segregation between split intein fragments can be used to manipulate the reactivities of different N- and C-intein pairs and thus engender reaction specificity on a system containing multiple split inteins.

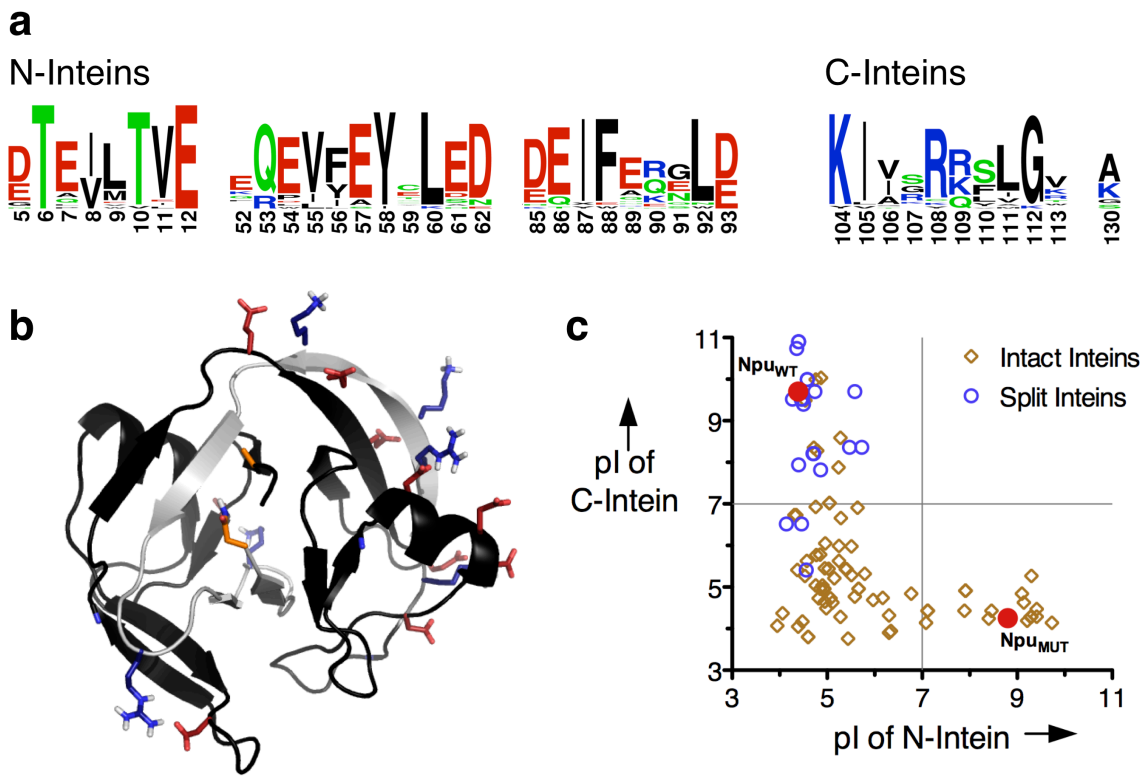


Figure 4.1. Charge segregation in split inteins. a. Sequence logo based on a sequence alignment of 24 naturally split inteins highlighting highly conserved charged residues. **b.** NMR structure of Npu$_{WT}$ (PDB code: 2KEQ) highlighting intermolecular electrostatic interactions. NpuN$_{WT}$ and NpuC$_{WT}$ are shown in black and gray, respectively. Acidic residues are shown in red, basic residues are shown in blue, and terminal catalytic residues are shown in orange. **c.** Comparison of calculated isoelectric points (pI) for the N- and C-terminal fragments of 24 naturally split inteins, the N- and C-terminal sequences of 76 intact inteins, and the mutant split intein (Npu$_{MUT}$) engineered in this study. Npu$_{WT}$ and Npu$_{MUT}$ are both indicated on the plot as red circles.

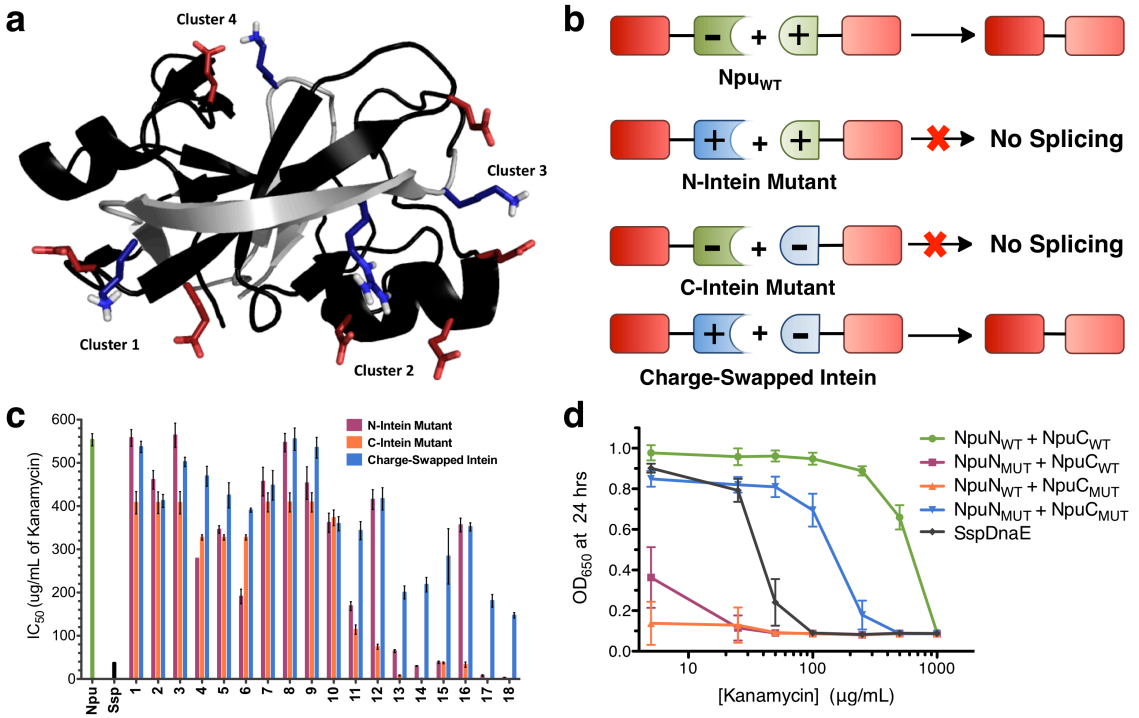


Figure 4.2. ***In vivo*** **mutational analysis of four ion clusters. a.** Rendering of the Npu structure highlighting the four ion clusters. **b.** Scheme depicting the expected specificity in charge-swapped inteins. **c.** IC_{50} values for all of the mutants tested in the intein activity-dependent kanamycin resistance assay. Mutants are defined in Table 4.1. **d.** Dose-response curves for Npu_{WT} and Npu_{MUT} fragment combinations in an *in vivo* intein activity-coupled kanamycin resistance assay. Mutation clusters in Npu_{MUT} are (E7K, K130E), (E52K, E54K, K113E), (D85K, E89R, R108E), and (E61K, E91K, K104E).

4.1. *In vivo* survey of charge-swapped mutants in Npu

To evaluate the role of specific ionic interactions for split intein splicing and specificity, we screened a series of charge-swapped Npu intein mutants and tested their splicing activity *in vivo*. Starting from the wild-type protomers, $NpuN_{WT}$ and $NpuC_{WT}$, mutation positions were chosen based on two criteria: i) residues should be involved in an intermolecular ion pair or triad, and ii) these residues should be moderately isolated from intramolecular ionic interactions and the active site. Based on these criteria, we identified four ion clusters comprised of one pair and three triads (Figure 4.2a). For each cluster and several combinations of clusters, three variants were generated: an N-intein mutant in

which native acidic residues were mutated to basic residues, a C-intein mutant in which basic residues were mutated to acidic residues, and a charge-swapped intein that combined all N- and C-intein mutations (Figure 4.2b). The activity of each set was analyzed using an *in vivo* splicing assay where kanamycin resistance in *E. coli* was dependent on intein splicing activity (Figure 4.2c and Table 4.1).[92] Of all combinations tested, the intein fragments in which all possible ion clusters were charged swapped (#18 in Table 4.1), referred to as $NpuN_{MUT}$ and $NpuC_{MUT}$, showed the least cross-reactivity with wild-type fragments *in vivo* (Figure 4.2d). Specifically, $NpuN_{WT}$ did not react with $NpuC_{MUT}$, $NpuN_{MUT}$ reacted minimally with $NpuC_{WT}$, and both mutant fragments displayed substantial splicing activity when combined *in vivo*. Importantly, the new charge-swapped intein, Npu_{MUT}, catalyzed protein splicing more efficiently than the widely used SspDnaE intein (Figure 4.2d).

Table 4.1. N- and C-intein mutation combinations analyzed *in vivo*.

	Ion Cluster 1			**Ion Cluster 2**			**Ion Cluster 3**			**Ion Cluster 4**	
	IntN		**IntC**	**IntN**		**IntC**	**IntN**		**IntC**	**IntN**	**IntC**
Intein	E52K	E54K	K113E	D85K	E89R	R108E	E61K	E91K	K104E	E7K	K130E
1	X		X								
2		X	X								
3	X	X	X								
4				X		X					
5					X	X					
6				X	X	X					
7							X		X		
8								X	X		
9							X	X	X		
10										X	X
11	X	X	X	X	X	X					
12	X	X	X				X	X	X		
13				X	X	X	X	X	X		
14	X	X	X	X	X	X	X	X	X		
15	X	X	X	X	X	X				X	X
16	X	X	X				X	X	X	X	X
17				X	X	X	X	X	X	X	X
18 [a]	X	X	X	X	X	X	X	X	X	X	X

[a] Intein mutant 18 refers to is Npu_{MUT} used throughout the rest of the chapter.

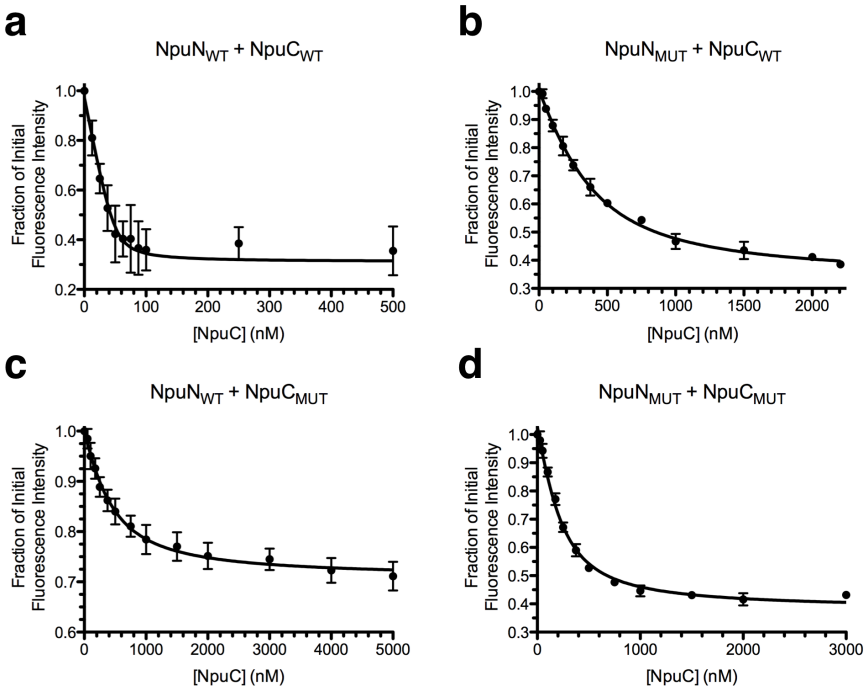


Figure 4.3. Fluorescence binding curves for Npu_{WT} and $NpuN_{MUT}$ fragments. a. Titration of $NpuC_{WT}$-SUMO against 50 nM Ub-$NpuN_{WT}$. **b.** Titration of $NpuC_{WT}$-SUMO against 300 nM Ub-$NpuN_{MUT}$. **c.** Titration of $NpuC_{MUT}$-SUMO against 300 nM Ub-$NpuN_{WT}$. **d.** Titration of $NpuC_{MUT}$-SUMO against 200 nM Ub-$NpuN_{MUT}$. In each case, NpuN has the C1A mutation. Samples were excited at 295 nm and emission was recorded at 345 nm. Curves were fit to a quadratic binding equation. (Error bars ± SD, n = 3)

4.2. The effect of ionic interactions on fragment binding affinity

To determine if the observed trend for *in vivo* splicing activities was a direct result of relative fragment binding affinities, we developed an *in vitro* binding assay. Using ubiquitin (Ub) and the small ubiquitin-like modifier (SUMO) as model N- and C-extein domains, we expressed and purified Ub-NpuN and NpuC-SUMO fusion proteins bearing wild-type or mutant intein sequences. The Ub-NpuN fusions contained a cysteine to alanine mutation (C1A) to prevent *trans*-splicing. To measure binding affinities, we took advantage of the single tryptophan residue (W47) in the Npu intein and measured a decrease in the intrinsic fluorescence of NpuN in the presence of increasing concentrations of NpuC-SUMO (Figure 4.3 and Table 4.2). As expected, the wild-type

fragments had a low-nanomolar binding affinity, consistent with previous measurements for the highly homologous SspDnaE intein.[50] The fully charged-swapped Npu_{MUT} N- and C-inteins associated with a 40-fold weaker affinity than the Npu_{WT} protomers (K_D = 118.4 nM), which may explain their slightly diminished splicing activity *in vivo*. Surprisingly, both wild-type/mutant hybrids had measurable binding affinities only two to three-fold weaker than $NpuN_{MUT}$ + $NpuC_{MUT}$, despite their extremely low or undetectable levels of splicing *in vivo*. These data indicated that while charge-swapping modulated intein fragment affinities with the expected trend, the magnitude of these energetic effects do not fully explain the splicing selectivity observed *in vivo*. Notably, we observed that the $NpuN_{WT}$ + $NpuC_{MUT}$ combination showed only a 29% decrease in N-intein fluorescence upon binding (Figure 4.3c), while all other fragment combinations showed a 60-70% decrease in fluorescence (Figure 4.3a,b,d). This anomalous fluorescence strongly suggests that the $NpuN_{WT}$ + $NpuC_{MUT}$ complex adopts a unique conformation which could explain its lack of splicing activity *in vivo*.

Table 4.2. *In vitro* characterization of Npu_{WT} and Npu_{MUT}.

Fragment pair		K_D (nM) [a]	Initial $t_{1/2}$ of *trans*-splicing (min) [a,b]		
NpuN	NpuC		1.0 μM	0.1 μM	0.05 μM
WT	WT	2.9 ± 1.8	0.9 ± 0.1	1.1 ± 0.2	2.4 ± 0.4
MUT	WT	211.6 ± 21.0	8.2 ± 0.4	71.0 ± 4.9	350.3 ± 41.2
WT	MUT	281.8 ± 51.5	n/a [c]	n/a [c]	n/a [c]
MUT	MUT	118.4 ± 11.4	6.5 ± 0.4	43.7 ± 2.1	103.0 ± 8.7

[a] Error values indicate the standard error from a global fit of three replicate data sets.
[b] Initial half-lives are calculated from a second-order rate constant.
[c] No detectable *trans*-splicing was observed for the $NpuN_{WT}$ + $NpuC_{MUT}$ combination.

4.3. *In vitro* splicing activity of mutant Npu fragments

We next examined the relative rates of *trans*-splicing for Npu_{WT}, Npu_{MUT}, and their N- and C-intein combinations *in vitro*. Splicing kinetics at 30 °C were measured by pair-wise mixing of Ub-NpuN and NpuC-SUMO fusions at equimolar ratios, monitoring

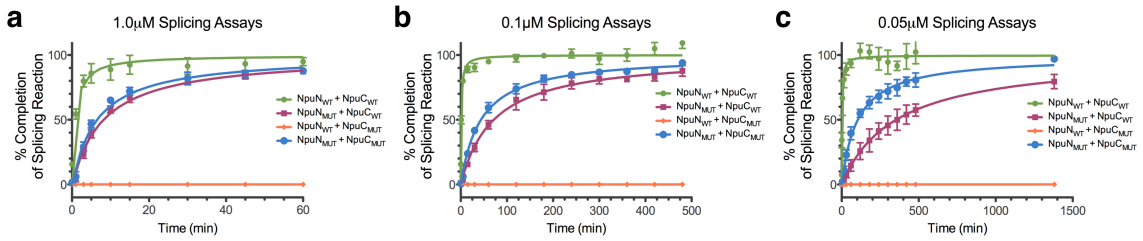


Figure 4.4. *In vitro* splicing kinetics for Npu_{WT} and Npu_{MUT} fragments. Normalized reaction progress curves at **a.** 1.0 μM, **b.** 0.1 μM, and **c.** 0.05 μM fit to a second order rate equation. $NpuN_{WT}$ + $NpuC_{WT}$ is green, $NpuN_{MUT}$ + $NpuC_{WT}$ is purple, $NpuN_{WT}$ + $NpuC_{MUT}$ is orange, and $NpuN_{MUT}$ + $NpuC_{MUT}$ is blue. (Error bars ± SD, n = 3)

the formation of the Ub-SUMO spliced product by gel electrophoresis (Figure 4.4 and Table 4.2). Npu_{WT} showed extremely rapid splicing ($t_{1/2}$ ~1 min) as previously observed.[89] As expected, given the low-nanomolar affinity for the wild-type fragments, this rapid rate of splicing was relatively independent of intein concentrations from 1.0 μM to 0.05 μM. Consistent with the *in vivo* splicing and *in vitro* binding experiments, Npu_{MUT} spliced slightly slower than Npu_{WT} at 1.0 μM ($t_{1/2}$ = 6.5 min), but its rate decreased roughly 10-fold at intein concentrations near its K_d. Surprisingly, the $NpuN_{MUT}$ + $NpuC_{WT}$ combination catalyzed splicing almost as quickly as $NpuN_{MUT}$ + $NpuC_{MUT}$ at high concentrations, but this difference in *trans*-splicing rates increased dramatically at lower concentrations, due to the weaker binding affinity for $NpuN_{MUT}$ + $NpuC_{WT}$. The $NpuN_{WT}$ + $NpuC_{MUT}$ combination showed no detectable *trans*-splicing at all three concentrations tested. Not only is this observation is consistent the *in vivo* results, but it supports the fluorescence data suggesting that this N- and C-intein bind to form a catalytically incompetent complex. The *in vitro* binding and splicing assays collectively demonstrate that charge complementation or repulsion can bias relative split intein binding affinities, which in turn bias relative splicing kinetics. These experiments also shed light on our *in vivo* splicing assays. Specifically, the degree of concentration

dependence on splicing rates observed *in vitro* indicates that the intein concentration *in vivo* is probably low nanomolar, at which level the *in vivo* and *in vitro* results would be consistent.

4.4. Orthogonal reactivity of wild-type and charge-swapped Npu

Given the differences in splicing kinetics observed *in vitro*, we envisioned that Npu_{WT} and Npu_{MUT} could be used simultaneously to carry out multiple *trans*-splicing reactions with kinetically controlled selectivity.[134,135] To test this, we developed an *in vitro* competition assay (Figure 4.5a) using our splicing assay fusion proteins and two additional fusions to the Npu_{WT} fragments bearing unique exteins, maltose binding protein (MBP) and enhanced green fluorescent protein (eGFP). As a control reaction, we mixed two $NpuN_{WT}$ fusions and two $NpuC_{WT}$ fusions bearing unique exteins at equimolar concentrations (0.5 μM) at 30 °C. The formation of all four possible products was monitored by western blotting against an HA epitope tag on the C-terminus of both C-exteins. As expected, all products formed to a similar extent in the control reaction (Figure 4.5b,c). When Ub-$NpuN_{MUT}$ and $NpuC_{MUT}$-SUMO were used in place of their wild-type counterparts, a clear bias in product formation was observed. The Npu_{WT} product (MBP-eGFP) formed rapidly, and the Npu_{MUT} product (Ub-SUMO) emerged slightly slower, consistent with our *in vitro* splicing assays (Figure 4.5d,e). These products formed almost exclusively, close to 50% each, while the $NpuN_{MUT}$ + $NpuC_{WT}$ product (Ub-eGFP) accounted for less than 5% of the total product formed, and the $NpuN_{WT}$ + $NpuC_{MUT}$ product (MBP-SUMO) was not observed (Figure 4.5e). These results were reproducible over a range of temperatures, from 25 °C to 37 °C, and at

higher ionic strength, indicating that the electrostatically driven selectivity between Npu_{WT} and Npu_{MUT} is extremely robust.

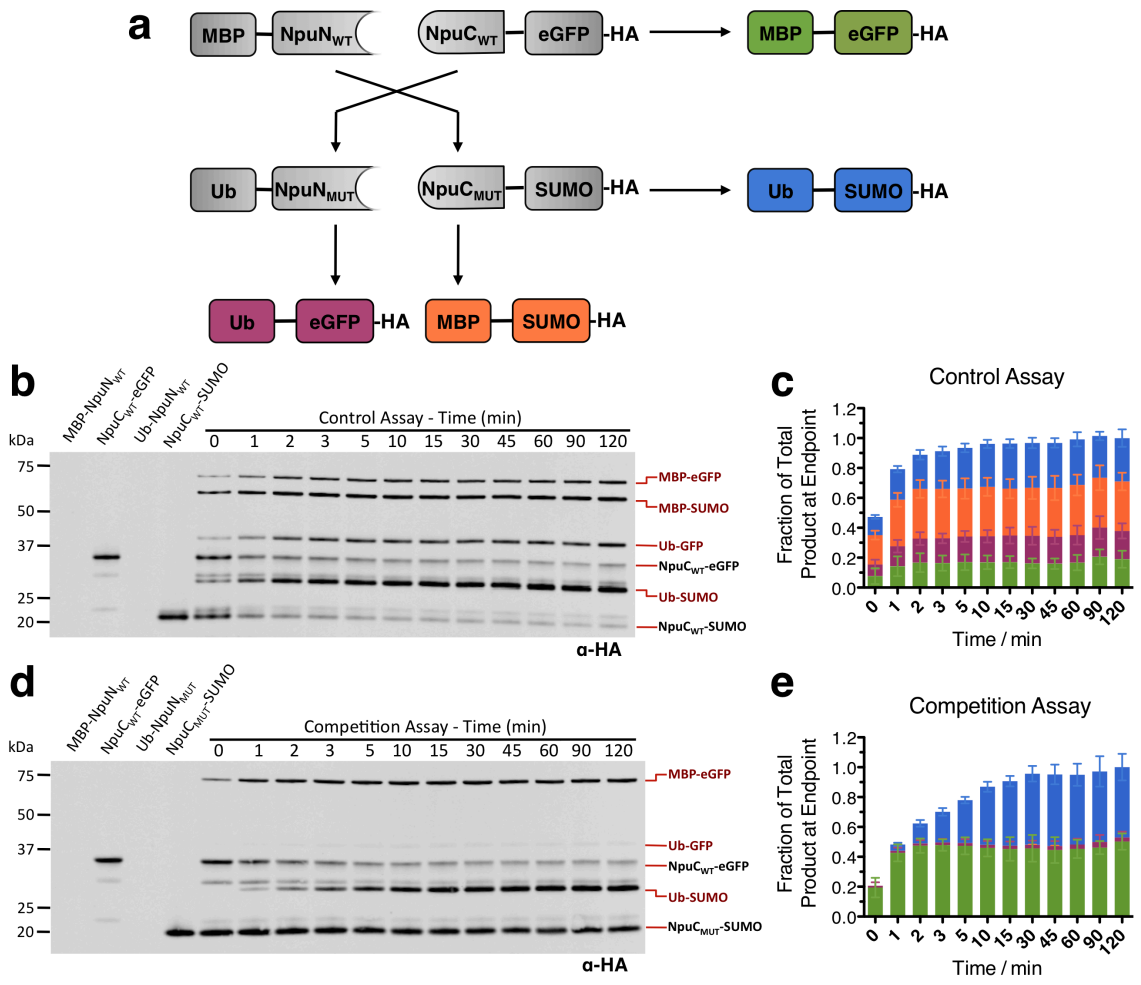


Figure 4.5. Competition splicing assays with Npu_{WT} and Npu_{MUT}. a. Scheme showing four possible splicing reactions in the competition assay. **b.** Western blot of the control reaction with wild-type intein fragments. **c.** Densitometric analysis of product bands in the control reaction. **d.** Western blot of the competition reaction with wild-type and mutant fragments. **c.** Densitometric analysis of product bands in the competition reaction.

4.5. Three-piece protein ligations using orthogonal split inteins

4.5.1. Ligation of a model system

Next, we explored the utility of these inteins for one-pot three-piece ligations of proteins (Figure 4.6a), as this type of ligation scheme could be useful for the segmental

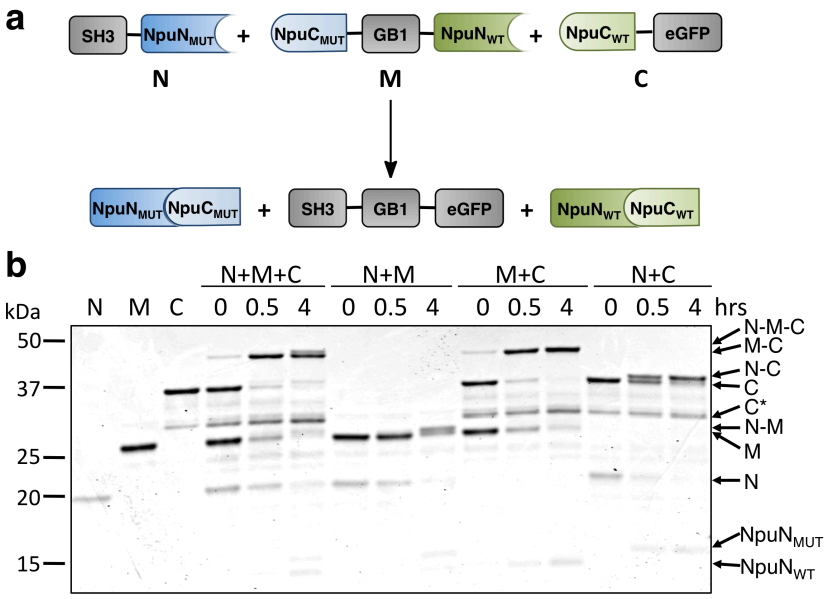


Figure 4.6. Three-piece ligation of a model system. a. Scheme depicting three-piece ligation of an SH3 domain, GB1, and eGFP using Npu_{WT} and Npu_{MUT}. **b.** SDS-PAGE analysis of two- and three-piece test reactions with all three fragments visualized by coomassie-staining. Note that the N-fragment stains poorly. C* refers to an impurity in the C-fragment preparation.

isotopic labeling of central regions of multi-domain proteins. Previously, orthogonal intein systems have been developed by combining naturally and artificially split inteins or using a wild-type and linearly permuted split intein.[50,100] While these systems could catalyze three-piece ligations, the prior showed low efficiency, and the latter could not be carried out in one pot. We envisioned that our Npu_{WT} and Npu_{MUT} pair could efficiently catalyze one-pot three-piece ligations, given the high degree of kinetic control seen in our competition assay. To test this, we designed a model system to ligate a Src-homology 3 (SH3) domain, domain B1 of protein G (GB1), and eGFP. The domains were fused to split intein sequences to generate an N-terminal fragment (N: SH3-$NpuN_{MUT}$), a middle fragment (M: $NpuC_{MUT}$-GB1-$NpuN_{WT}$), and a C-terminal fragment (C: $NpuC_{WT}$-eGFP). Importantly, we designed the middle domain flanked by the N- and C-inteins that do not

react in *trans* since this would preclude GB1 cyclization or oligomerization (Figure 4.6a). To test our three-piece ligation system, we mixed the N, M, and C fragments in pairs or all together with slight excess of the middle fragment. Analysis of the reaction mixtures by gel electrophoresis (Figure 4.6b) indicated that the reaction between $NpuN_{WT}$ and $NpuC_{WT}$ (M+C) occurred extremely fast to yield the desired intermediate product. In the three-piece reaction, this intermediate was more slowly converted to the full-length three-piece-ligated product, SH3-GB1-eGFP, upon reaction with N. Importantly, while the N+C reaction could proceed in the absence of the middle piece, the formation of this undesired product (SH3-eGFP) was suppressed in the three-piece reaction, consistent with the kinetically controlled *trans*-splicing paradigm.

4.5.2. Assembly of poly(ADP-ribose)-polymerase 1 from three pieces

Next, we sought to apply this three-piece ligation technology towards the semi-synthesis of human poly(ADP-ribose) polymerase 1 (PARP1). This ~115 kDa enzyme, which catalyzes the transfer of ADP-ribose from nicotinamide adenine dinucleotide (NAD) onto protein side chains in the form of monomers or polymers, is involved DNA-damage pathways and is a promising target for chemotherapeutics.[136,137] Given its large size and its capacity to automodify itself,[138] full-length PARP1 is not readily isolable by over-expression in *E. coli*.[139,140] We envisioned that our orthogonal inteins could be used to generate full-length PARP1 by expression of its fragments separately in *E. coli* followed by *in vitro* three-piece ligation. Based on known domain boundaries,[141,142] we chose two ligation sites in putatively flexible regions that separated PARP1 into three fragments: the N-terminal zinc finger DNA-binding domains (PARP1-N), the central

dimerization and automodification domains (PARP1-M), and the C-terminal catalytic domain (PARP1-C) (Figure 4.7a). Importantly, each of these functional domains is required for activation and regulation of PARP1 catalysis,[143] and thus their efficient and accurate assembly is a rigorous test of our three-piece ligation system.

Due to a lack of cysteine residues near the desired splice junctions, we introduced two modest mutations to allow for split intein catalysis, S364C and T656C (Figure 4.7a), and fused the PARP1 fragments to our split inteins. In addition, we designed a traceless tagging strategy in which His_6-tags were placed at the free termini of every intein sequence. These tags could be used not only to facilitate enrichment of the proteins after over-expression but also to trap unreacted starting materials, intermediates, and spliced intein fragments after a three-piece ligation. To generate full-length PARP1, we expressed the three intein-fusion fragments separately in *E. coli* and enriched the proteins over nickel columns. The semi-pure proteins were mixed and allowed to react at room temperature for 21 hours. Following the reaction by western blotting against PARP1-N and PARP1-C, we observed the formation of both reaction intermediates as well as full-length PARP1 (Figure 4.7b). The reaction mixture was passed through a nickel affinity column to trap residual starting materials, intermediates, and free intein fragments. The flow-through, containing the full-length product, was then purified by size exclusion chromatography to yield pure, three-piece-ligated PARP1 (Figure 4.7b). Mass spectrometric analysis of tryptic fragments confirmed the identity of the purified product (Figure 4.7c), and peptides spanning the ligation junctions were further analyzed by MS/MS to confirm their sequences (Figure 4.7d,e).

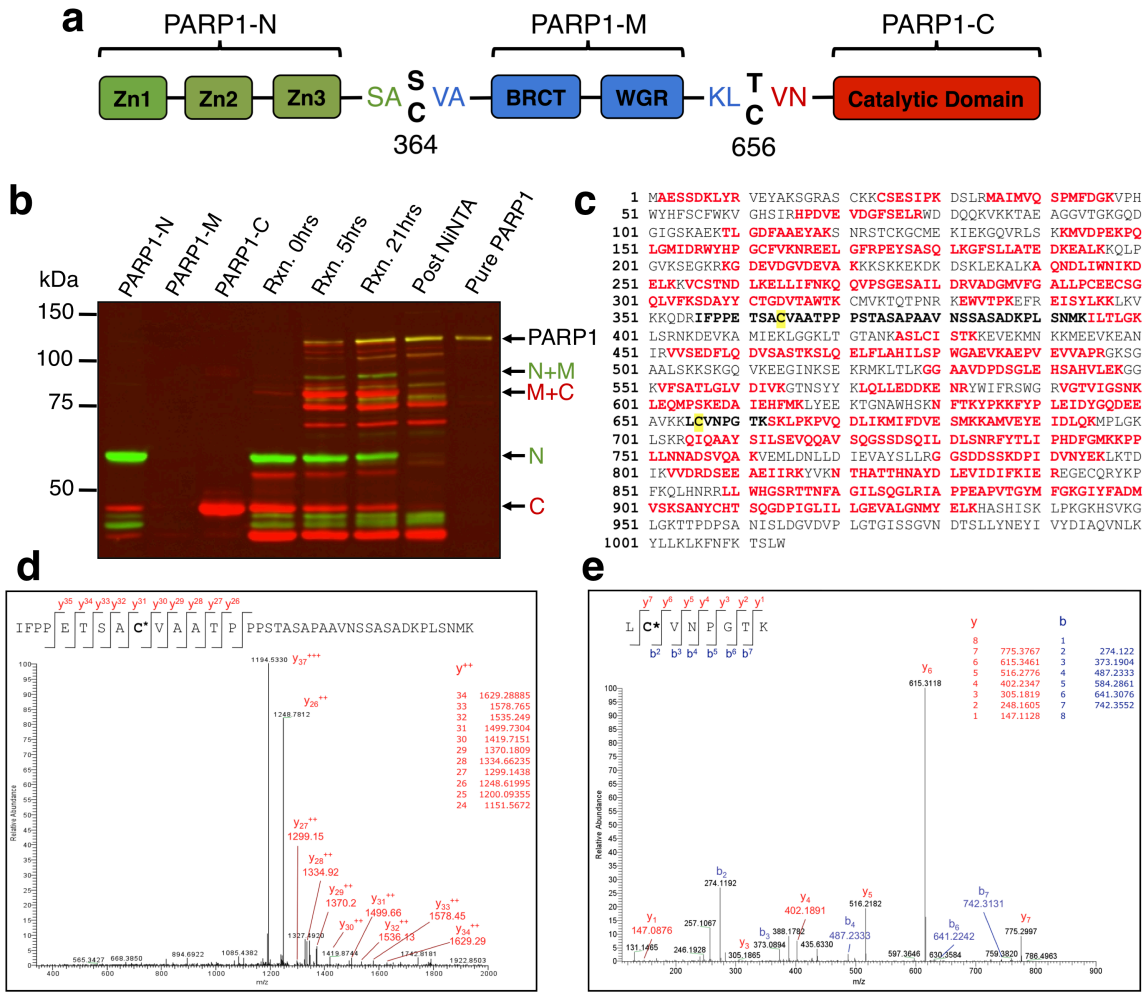


Figure 4.7. Three-piece ligation of human PARP1. a. Domain organization and splice junctions in PARP1. **b.** Western blot of PARP1 three-piece ligation and purification. Color scheme: green bands (α-PARP1-N), red bands (α-PARP1-C), and yellow bands (α-PARP1-N + α-PARP1-C). **c.** Sequence coverage of tryptic peptides identified by MS. Observed peptides are shown in red. Sequences in bold black correspond to tryptic peptides spanning the ligation junctions. MS/MS analysis of the peptides surrounding the **d.** first and **e.** second ligation sites confirming their sequences.

To determine whether our semi-synthetic PARP1 was active, we conducted ADP-ribosylation assays using biotinylated-NAD and streptavidin blotting. We observed that our three-piece ligated PARP1 could catalyze automodification (Figure 4.8a). Importantly, this activity required activated DNA, could be inhibited by benzamide, and was stimulated by histone octamers, all of which are hallmarks of full-length PARP1

activity.[144,145] Furthermore, our enzyme could catalyze the formation of poly(ADP-ribose) chains on histones, as previously reported for endogenous PARP1 (Figure 4.8b).[146] These enzymatic data unequivocally demonstrate that our three-piece ligated PARP1 is full-length and properly folded, as the DNA-dependent activity of this enzyme requires allosteric communication between all three segments. Consistent with this requirement, the PARP1-C fragment, bearing only the catalytic domain of PARP1, did not catalyze poly(ADP-ribosylation) of histones (Figure 4.8c).

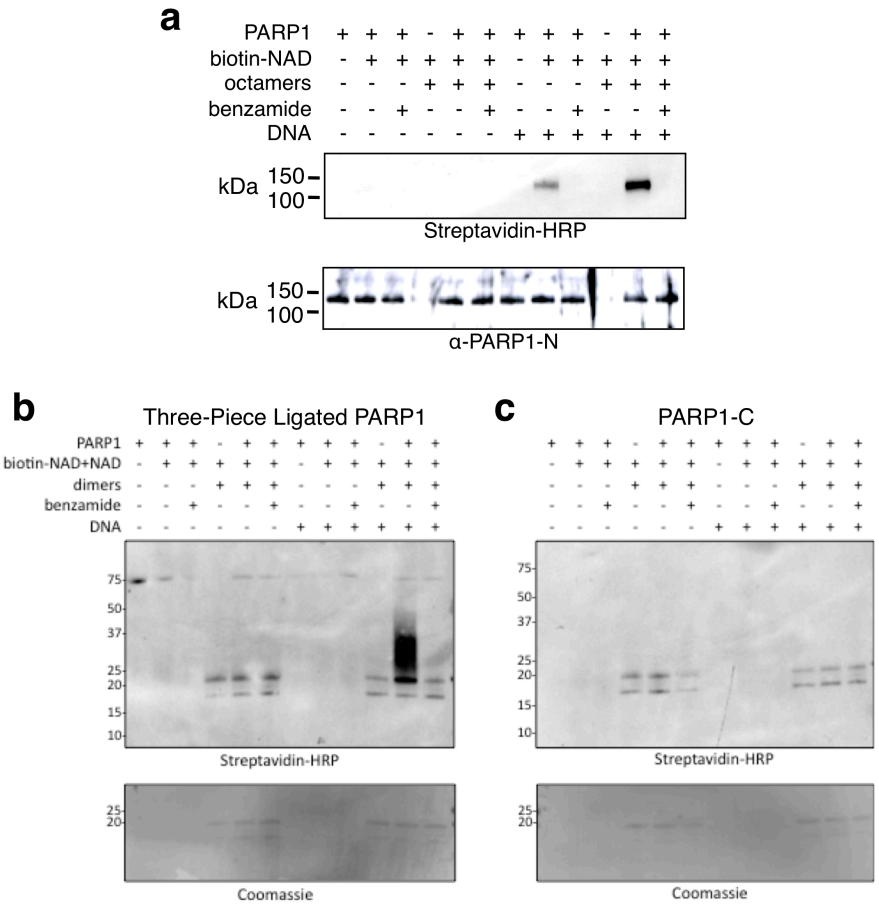


Figure 4.8. Catalytic activity of three-piece ligated PARP1. a. Biotinylated-NAD blot showing automodification of PARP1 (streptavidin-HRP) and PARP1 blot loading control (α-PARP1-N). **b.** Biotinylated-NAD blot showing poly(ADP)-ribosylation of histones with ligated PARP1. **c.** Biotinylated-NAD blot showing lack of poly(ADP)-ribosylation on histones with PARP1-C. The coomassie-stained membranes are shown below as histone loading controls.

4.6. Summary and conclusions

In this study, we probed the role of intermolecular ion clusters for fragment assembly and splicing in the Npu_{WT} split intein both *in vivo* and *in vitro*. Our mutagenesis experiments demonstrate that charge segregation and specific intermolecular electrostatic interactions are important for split intein fragment association and thus splicing. Through these experiments, we rationally designed a new split intein, Npu_{MUT}, which displays low cross-reactivity with Npu_{WT}. These orthogonal inteins were used to generate the large, full-length, active mammalian protein, PARP1, through a one-pot three-piece ligation. Collectively, our results demonstrate that electrostatic interactions can engender kinetic control and thus specificity in a complex biochemical system. Furthermore, these results provide the first insights into the molecular requirements for fragment complementation in naturally split inteins. In the next chapter, the mechanism of split intein assembly is examined in greater detail, highlighting not only the importance of the electrostatic interactions described herein, but also the unique structural properties of split intein fragments before they encounter one another.

Chapter 5: The mechanism of split intein coupled binding and folding*

There are two types of naturally occurring inteins: highly abundant contiguous inteins (Figure 5.1a) and rarer split inteins (Figure 5.1b). The latter carry out protein splicing *in trans*, where the biochemistry is preceded by the spontaneous association of two separately transcribed and translated fragments (N- and C-inteins). Although naturally split inteins evolved from their contiguous counterparts and share the same fold,[17,19] it is notable that many artificially split versions of contiguous inteins cannot spontaneously assemble and splice.[65,82,147] The intein fold has a complex interwoven topology (Figure 5.1c),[39] and split inteins must bear unique features that allow them to access this topology *in trans*. Given that inteins splice essential proteins, there would have been strong selective pressure to optimize fragment complementation in early split inteins. Indeed, this evolution is evident in modern split inteins, which associate rapidly and tightly.[50,98,148]

The largest known class of split inteins is the cyanobacterial DnaE intein family.[19] In their natural context, these inteins assemble the catalytic subunit of DNA polymerase III (DnaE). Several members of this family carry out protein splicing with remarkable efficiency, in tens of seconds rather than hours as is the case for many other inteins, making them ideal tools for protein engineering (discussed in Chapter 2).[89,123] Of these

---

* The work described in this chapter was carried out in close collaboration with Ertan Eryilmaz in Professor David Cowburn's lab at Albert Einstein College of Medicine and was published in the following paper:

Shah, N. H., Eryilmaz, E., Cowburn, D., and Muir, T. W., Naturally split inteins assemble through a "capture and collapse" mechanism. *J. Am. Chem. Soc.* **135**, 18673-18681 (2013).

ultrafast *trans*-splicing domains, the DnaE intein from *Nostoc punctiforme* (Npu) is the best-characterized member, and it is rapidly becoming the intein of choice for new protein chemistry technologies.[86,93,124,125] Currently, little is known about the assembly of Npu or other split DnaE inteins. We previously demonstrated that the Npu fragments associate with low nanomolar affinity (section 4.2),[98] similar to orthologous split intein fragments from *Synechocystis sp.* PCC6803 (Ssp).[50,148] Atomic force microscopy measurements on three other split DnaE inteins suggest that this tight binding is a conserved property in this family.[149] Furthermore, Ssp binding is known to be extraordinarily fast and ionic strength dependent.[50] Recently, it was proposed that both Ssp fragments undergo a disorder-to-order transition upon binding, however the molecular determinants and mechanism of this transition were not established.[148] Based on sequence analyses, it is clear that most split inteins differ from contiguous ones by the presence of significant charge segregation between the N- and C-inteins (Figure 4.1), and indeed, mutating critical ion pairs and triads at the fragment interface weakens fragment affinity (discussed in Chapter 4).[98]

Here we elucidate the mechanism of split intein assembly, using Npu as a relevant model system. We show that split inteins fragments interact through a multi-phase process initiated by electrostatic interactions between extended regions of both fragments (capture) followed by compaction and stabilization of the initially disordered segments onto a pre-folded region of the N-intein (collapse). Given the emerging role of split inteins as powerful tools for *in vitro* and *in vivo* protein engineering and semisynthesis,[150] this biophysical mechanism should aid in the development and further improvement of split intein-based technologies.

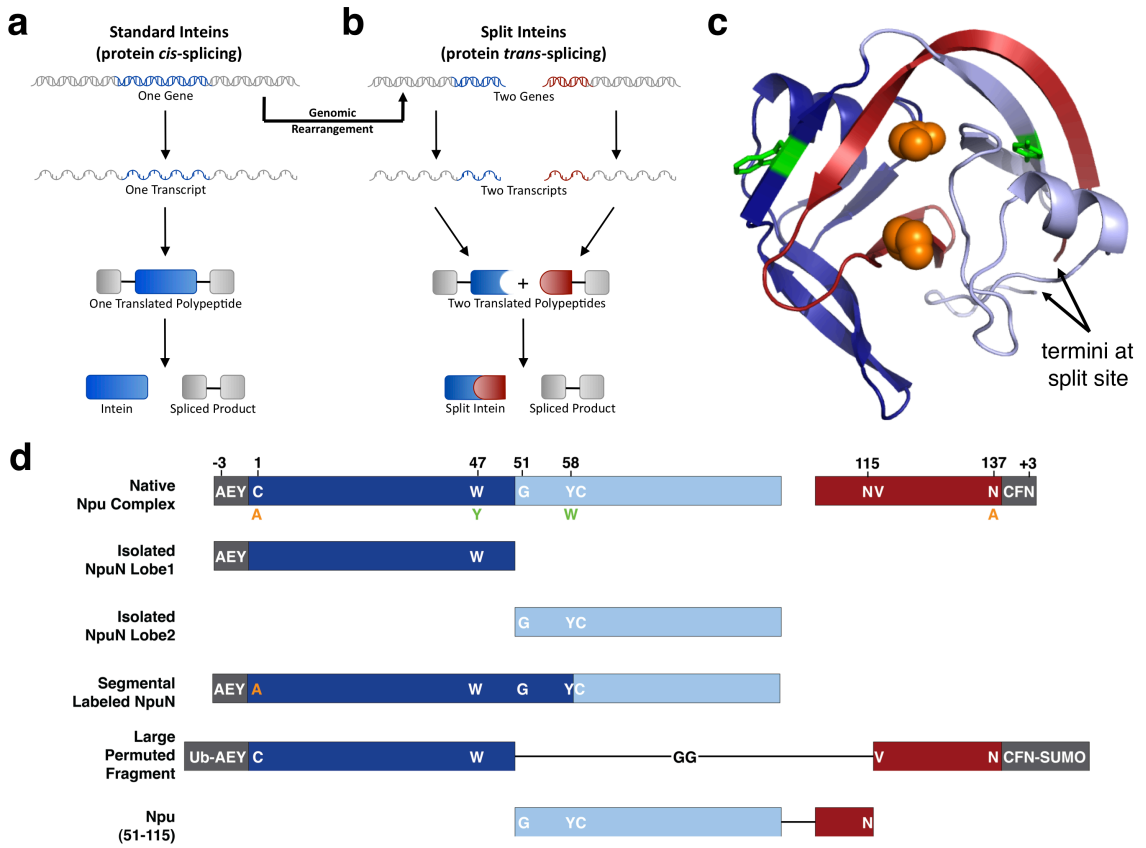


Figure 5.1. The origin and structure of split inteins. a. Scheme depicting protein *cis*-splicing with contiguous inteins. **b**. Scheme depicting protein *trans*-splicing with split inteins. **c.** Rendering of the Npu intein structure, highlighting relevant NpuN lobes ($NpuN_1$ in dark blue and $NpuN_2$ in light blue) and NpuC in red. Termini to which exteins would be attached are shown as orange spheres, and sites of intrinsic fluorophores used in this study (W47 and Y58W) are shown as green sticks; rendering based on PDB 2KEQ.[39] **d.** Scheme of constructs used in this study. Coloring of fragments and lobes is analogous to panel c. Key residues are labeled and mutations of those residues used in this study are indicated.

5.1. Npu fragments undergo dramatic conformational changes upon association

To elucidate the binding mechanism of Npu and related inteins, we initially sought to characterize the structures of the 102 residue N-intein (NpuN) and 35 residue C-intein (NpuC) in isolation. Both proteins were generated and purified as previously described (in Chapter 3),[151] bearing short tripeptide exteins that should not contribute significantly to the spectroscopic or chromatographic signals being observed (Figure

5.1d). These constructs also contained a C1A mutation in the N-intein and N137A mutation in the C-intein to prevent protein splicing during analyses of the split intein complex.[151] Circular dichroism spectra of the fragments and an equimolar mixture suggested low α and β secondary structure content in the isolated fragments and a dramatic structural change upon binding (Figure 5.2a). Consistent with these data, NpuC was extremely susceptible to degradation by thermolysin, whereas the complex was effectively resistant to proteolysis (Figure 5.3a,b). NpuN showed an intermediate rate of degradation, suggesting a partially compact structure (Figure 5.3c). Remarkably, the N-intein eluted earlier than the complex on a size exclusion chromatography (SEC) column, indicating that it must collapse from an extended state upon binding NpuC (Figure 5.2b). To rule out the possibility of dimerization, we performed multi-angle light scattering (MALS) on-line with SEC and confirmed that NpuN is a monomer in solution (Figure 5.2b).

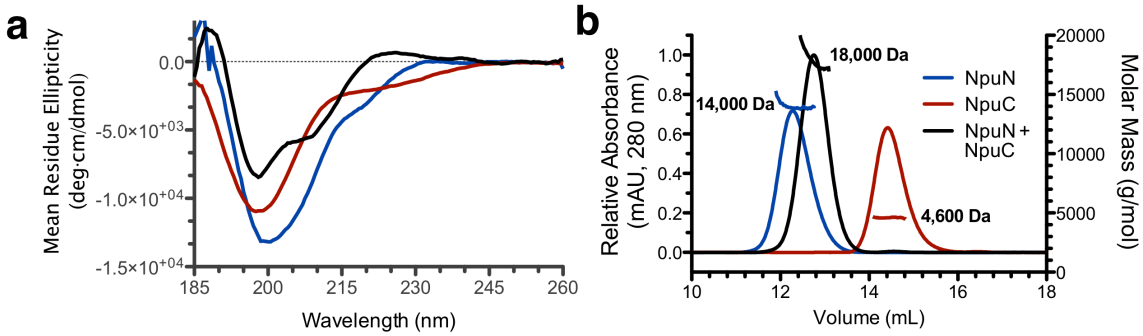


Figure 5.2. Circular dichroism and SEC-MALS of split intein fragments. a. Circular dichroism spectra were measured at 25 °C in a pH 7.2 buffer containing 25 mM sodium phosphates, 50 mM NaF, and 1 mM DTT. [NpuN] = 2.5 μM, [NpuC] = 12.5 μM, [Complex] = 2.5 μM. Cuvette pathlength = 1 mm. **b.** SEC-MALS of NpuN (blue, 25 μM, expected MW = 12188 Da), NpuC (red, 400 μM, expected MW = 4443 Da), and their complex (black, 25 μM, expected MW = 16631 Da) carried out on an S75 10/300 column in pH 6.5 buffer containing 25 mM phosphate, 100 mM NaCl, and 1 mM DTT.

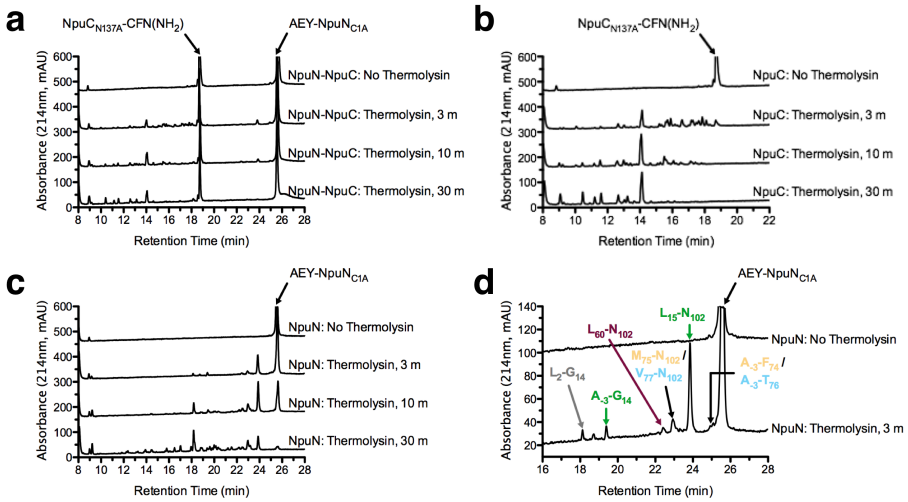


Figure 5.3. Limited proteolysis by thermolysin. Each substrate (12.5 μM, 0.2 mg/mL for the Npu complex) was incubated at 30 °C in the presence of 2 μg/mL of thermolysin in a pH 7.4 buffer containing 50 mM Tris HCl, 100 mM NaCl, 2 mM $MgSO_4$, 2 mM $CaCl_2$, and 1 mM DTT. **a.** Proteolysis of a 1:1 mixture of NpuN and NpuC. **b.** Proteolysis of NpuC alone. **c.** Proteolysis of NpuN alone. **d.** Annotation of the "single-cut" products from NpuN proteolysis identified by mass spectrometry (one site in $NpuN_1$ is labeled in green and three sites in $NpuN_2$ are labeled in yellow, blue, and purple). Note that $NpuN_1$ has 14 predicted cleavage sites out of 53 residues, $NpuN_2$ has 13 out of 52 residues, and NpuC has 14 out of 39 residues.

Additional insight into fragment structure was gained from nuclear magnetic resonance (NMR) spectroscopy. The ^1H-^{15}N HSQC spectrum of uniformly ^{15}N-labeled NpuC showed sharp amide N-H resonances with poor dispersion along the ^1H axis, indicative of an unfolded protein (Figure 5.4a, red). Notably, a similar spectrum was previously observed for the homologous Ssp C-intein, suggesting conserved structural properties among split inteins, even in the unbound state.[148] Upon addition of unlabeled NpuN, the signal dispersion increased dramatically, consistent with NpuC being part of a well-folded protein domain (Figure 5.4a, black). In the bound and unbound states, backbone resonances were readily assigned to aid in further analyses (Figure 5.4c). NMR relaxation experiments (R_1, R_2, and heteronuclear NOE measurements)[152] on isolated NpuC confirmed the high degree of flexibility throughout the chain with no detectable

residual secondary structure (Figure 5.4d-f). In contrast to NpuC, the HSQC spectrum of isolated NpuN showed moderate signal dispersion, however only a fraction of the expected resonances were visible, and the observed signals varied greatly in relative intensity (Figure 5.4b, blue). These data depict an N-intein that is highly dynamic but at least partially folded, consistent with both the expanded dimensions observed by SEC and the intermediate protection from thermolysin proteolysis. While these experiments provide a general picture of NpuN, the nature of the NMR data precluded backbone resonance assignments and thus sequence-specific information.

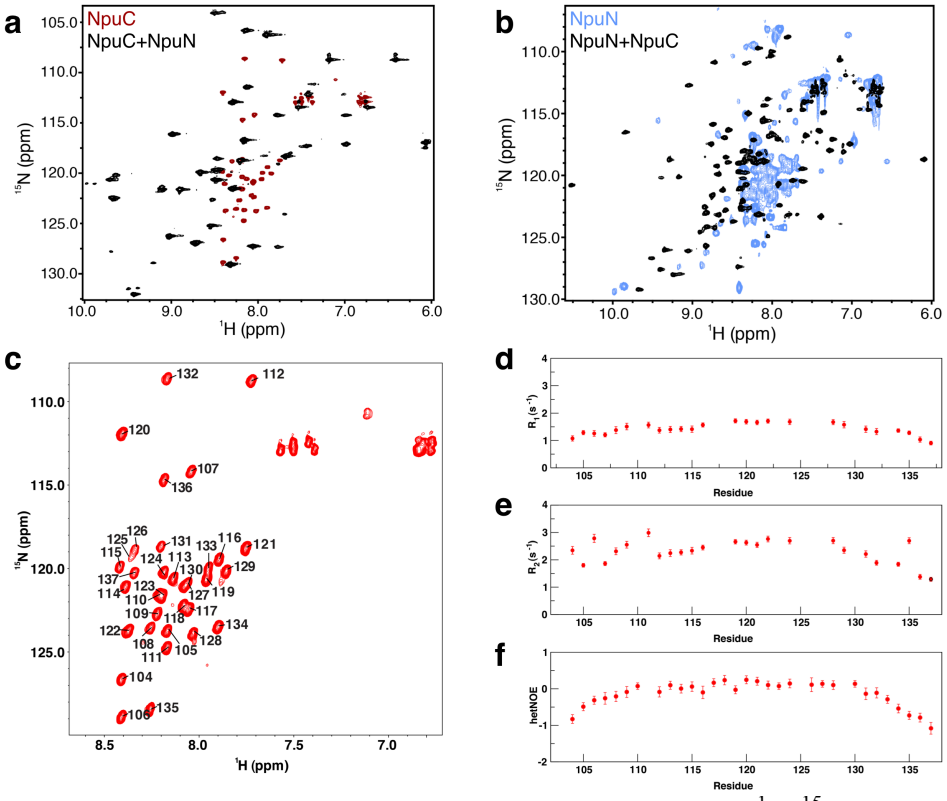


Figure 5.4. NMR characterization of split intein fragments. a. ^{1}H-^{15}N HSQC spectra of NpuC alone (red, 250 μM, 600 MHz) and in complex with unlabeled NpuN (black, 250 μM, 800 MHz). **b.** ^{1}H-^{15}N HSQC spectra of NpuN alone (blue, 250 μM, 500 MHz) and in complex with unlabeled NpuC (black, 250 μM, 800 MHz). **c.** ^{1}H-^{15}N HSQC spectrum of isolated NpuC, showing backbone NH assignments. Relaxation experiments for NpuC alone: **d.** R_1 relaxation rates, **e.** R_2 relaxation rates, and **g.** heteronuclear NOE values (^{1}H-^{15}N) for NpuC alone. NMR spectra were acquired at 25 °C in a pH 6.5 buffer containing 25 mM phosphates, 100 mM NaCl, and 1 mM DTT.

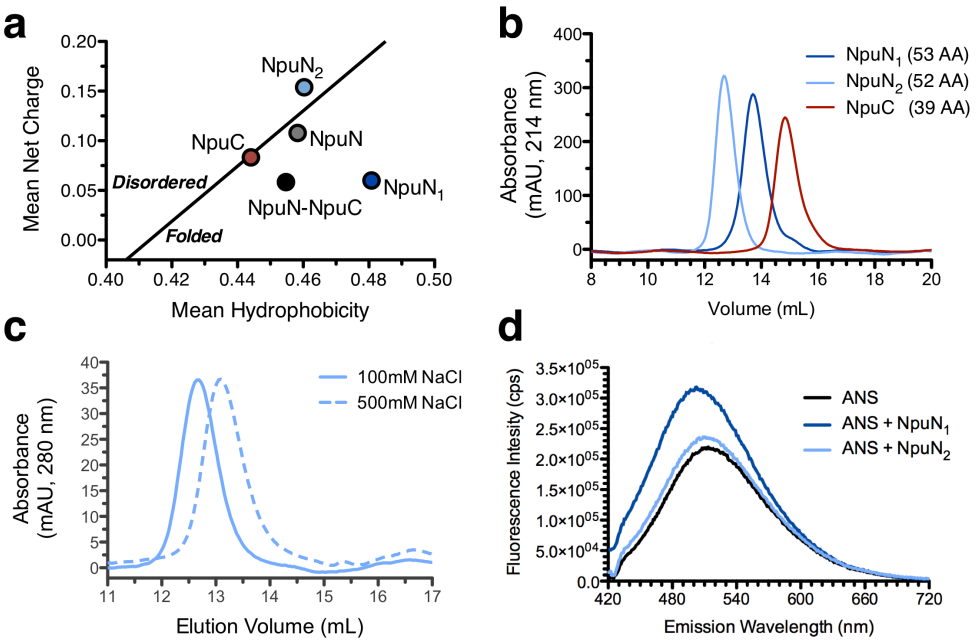


Figure 5.5. Comparison of isolated NpuN lobes. a. Charge-hydrophobicity plot comparing Npu fragments, complex, and N-intein lobes. Mean hydrophobicity (H) is calculated on a normalized Kyte-Doolittle scale, and mean net charge (R) is the absolute value. The solid line delineating disordered and folded proteins is empirically defined as $R = 2.785 \cdot H - 1.151$.[3] **b.** SEC of $NpuN_1$ (dark blue, 7 μM), $NpuN_2$ (light blue, 7 μM) and NpuC (red, 7 μM). **c.** SEC analysis of $NpuN_2$ compactness at different ionic strengths. **d.** Fluorescence of 4 μM ANS alone or in the presence of 4 μM $NpuN_1$ or $NpuN_2$. All experiments were done in a pH 6.5 buffer containing 25 mM phosphates, 100 mM NaCl, and 1 mM DTT except where indicated (panel c, dashed line).

5.2. NpuN is comprised of two structurally distinct lobes

We previously demonstrated that Npu fragment binding is dependent on electrostatic interactions between the highly acidic NpuN (calculated pI = 4.4) and highly basic NpuC (calculated pI = 9.7).[98] The majority of the relevant anionic residues on NpuN are concentrated on the second half of the sequence, which is spatially distinct from the first half of the sequence in the final intein complex (Figure 5.1c, light and dark blue regions, respectively). Further analysis of the sequence composition of NpuN indicated that the first lobe ($NpuN_1$, residues 1-50) has a high density of hydrophobic residues and low net charge per residue, consistent with a well-folded protein (Figure

5.5a). By contrast, the second lobe ($NpuN_2$, residues 51-102) has a high net charge per residue and lower hydrophobicity, a hallmark of intrinsically disordered proteins (IDPs, Figure 5.5a).[3] Indeed, despite their virtually identical molecular weights (ΔMW = 2 Da) and their symmetry-related structures in the final complex (Figure 5.1c), isolated $NpuN_1$ eluted later than $NpuN_2$ on a SEC column, suggesting a more compact structure for this lobe (Figures 5.5b). The more expanded lobe, $NpuN_2$, could be compacted in the presence of high salt, a known property of IDPs (Figure 5.5c).[153] Additionally, $NpuN_1$, but not $NpuN_2$, bound the hydrophobic dye 1-anilino-8-naphthalene sulfonate (ANS), suggesting that this compact lobe has molten globule-like properties with solvent-exposed hydrophobic surfaces (Figure 5.5d).[154] Unlike $NpuN_2$, the interface between $NpuN_1$ and NpuC in the final split intein complex is dominated by hydrophobic interactions (Figure 5.6), which presumably satisfy these exposed patches.

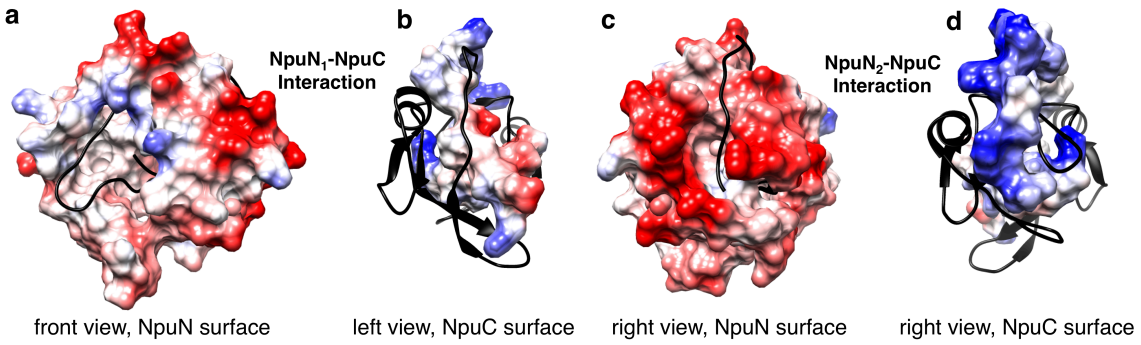


Figure 5.6. Electrostatic surface representations of Npu. The color scheme for these representation is red for negative charge, white for neutral, and blue for positive charge. All perspectives are given relative to panel a, which is the same perspective as Figure 5.1c. The renderings are based on PDB 2KEQ. **a.** and **b.** highlight complementary hydrophobic surfaces on either fragment at the $NpuN_1$-NpuC interface **c.** and **d.** highlight complementary electrostatic surfaces on either fragment at the $NpuN_2$-NpuC interface. Representations were made using UCSF Chimera.[155]

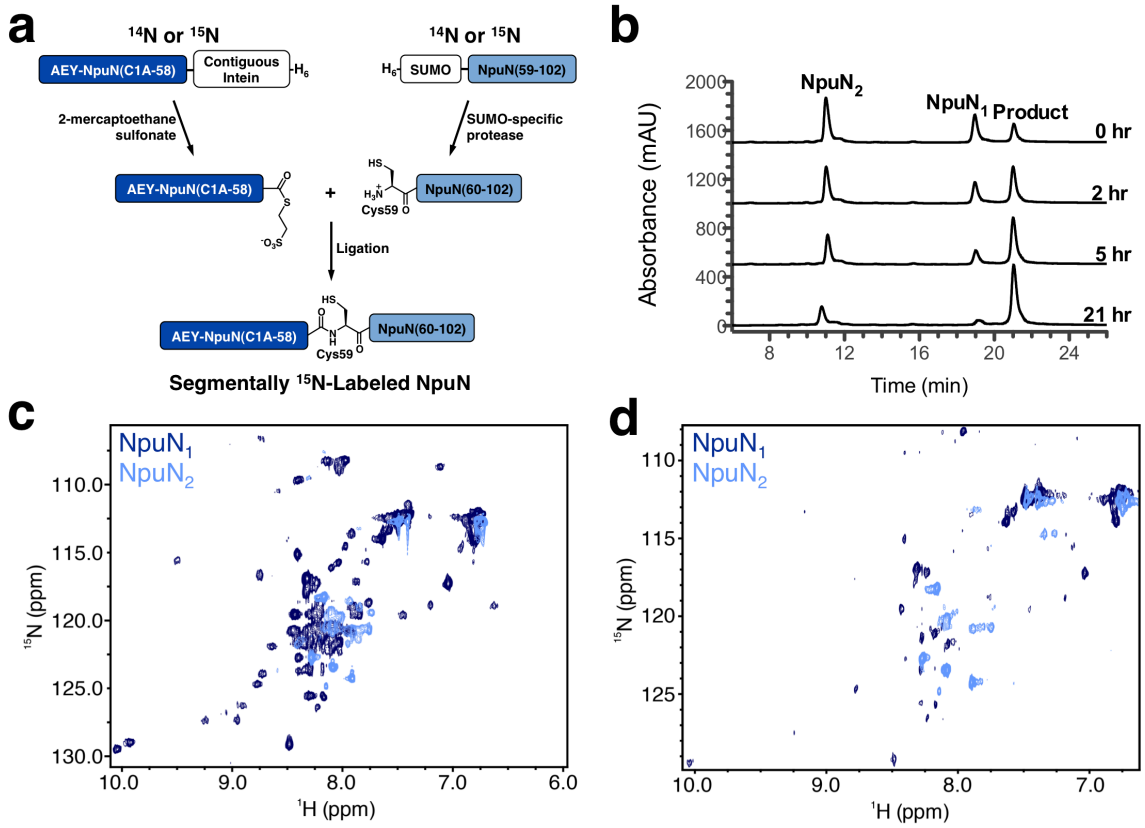


Figure 5.7. Segmental labeling of NpuN lobes. a. Scheme depicting the segmental isotopic labeling of NpuN using Expressed Protein Ligation. **b.** RP-HPLC analysis of the segmental labeling reaction progress. ^{1}H-^{15}N HSQC spectra of NpuN segmentally ^{15}N-labeled on $NpuN_1$ (dark blue, 100 μM) and $NpuN_2$ (light blue, 100 μM) collected at **c.** 500 MHz and **d.** 800 MHz. NMR data was acquired at 25 °C in a pH 6.5 buffer containing 25 mM phosphates, 100 mM NaCl, and 1 mM DTT.

We next asked if the different structural properties observed in the isolated NpuN lobes are retained in the context of full-length NpuN. To address this, we identified products from the limited proteolysis of NpuN by mass spectrometry. Early in the proteolysis reaction, we identified several pairs of products consistent with single proteolytic events (Figure 5.3d). As expected, in full-length NpuN, the second lobe was more susceptible to proteolysis than the first, supporting the notion that it is less compact. To more definitively assess the relative lobe structures, we employed Expressed Protein Ligation (EPL)[64] to generate full-length NpuN with segmental isotope labeling either on

$NpuN_1$ or $NpuN_2$. Residue 59 in NpuN, eight residues downstream of the $NpuN_1$-$NpuN_2$ junction, is a native cysteine and provides a reactive handle for ligation of the lobes while retaining the wild-type NpuN sequence (Figures 5.1d and 5.7a,b). This ligation approach allows for the selective observation of either $NpuN_1$ or $NpuN_2$ resonances in the context of full-length NpuN. The ^{1}H-^{15}N HSQC spectrum of the $NpuN_1$-labeled sample showed markedly greater proton chemical shift dispersion than that of the $NpuN_2$-labeled sample (Figure 5.7c). These spectra confirm the bipartite structure of the N-intein, with an at least partially folded $NpuN_1$ lobe and a disordered $NpuN_2$ lobe. Interestingly, many of the $NpuN_1$ resonances that were visible on a 500 MHz spectrometer were not visible at 800 MHz due to chemical exchange line-broadening, whereas $NpuN_2$ resonances were less affected by the field strength of the spectrometer (Figure 5.7d). Thus, the two connected lobes have different chemical exchange rates indicating internal dynamics on different time scales,[156] consistent with their distinct states of compaction. Importantly, both segmentally labeled N-inteins could bind NpuC to yield correctly folded, functional complexes (Figure 5.8).

5.3. The $NpuN_2$-NpuC interaction represents a binding intermediate

Given their distinct structures, we hypothesized that $NpuN_1$ and $NpuN_2$ play unique roles in engaging NpuC. To test this, we analyzed mixtures of each isolated lobe with NpuC by SEC-MALS at concentrations ranging from 7 μM to 50 μM. No interaction was observed between the two N-intein lobes or between $NpuN_1$ and NpuC at any concentration tested (Figure 5.9a,b). Remarkably, a 1:1 complex between $NpuN_2$ and NpuC was unambiguously observed by SEC-MALS (Figure 5.9c,d). It is noteworthy that

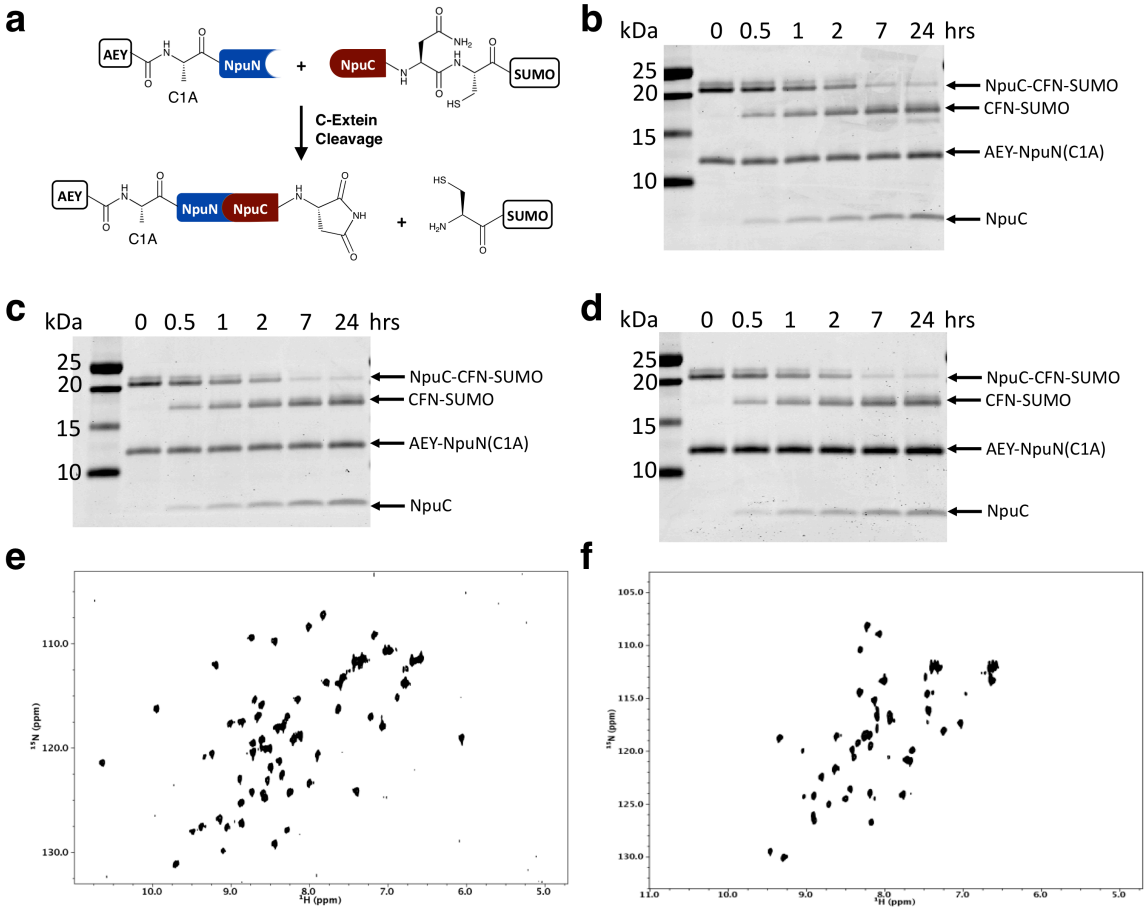


Figure 5.8. Control experiments with segmentally labeled NpuN. a. Scheme depicting control C-extein cleavage reaction to analyze NpuN activity with a C1A mutation. SDS-PAGE analysis of C-extein cleavage with a **b.** natively prepared NpuN sample, **c.** NpuN with segmental isotope labeling on the first lobe, and **d.** segmental isotope labeling on the second lobe. ^{1}H-^{15}N HSQC spectra of **e.** NpuN$_1$-labeled N-intein and **f.** NpuN$_2$-labeled N-intein in complex with unlabeled NpuC. Spectra were recorded on a 800 MHz spectrometer. All NMR experiments were carried out with 100 μM protein at 25 °C in a pH 6.5 buffer containing 25 mM sodium phosphates, 100 mM NaCl, and 1 mM DTT.

this complex eluted later than NpuN$_2$ alone, indicating that this lobe is compacted upon binding NpuC. Furthermore, binding was weakened by increasing the ionic strength, demonstrating that there is an electrostatic component to this interaction (Figure 5.9e,f).

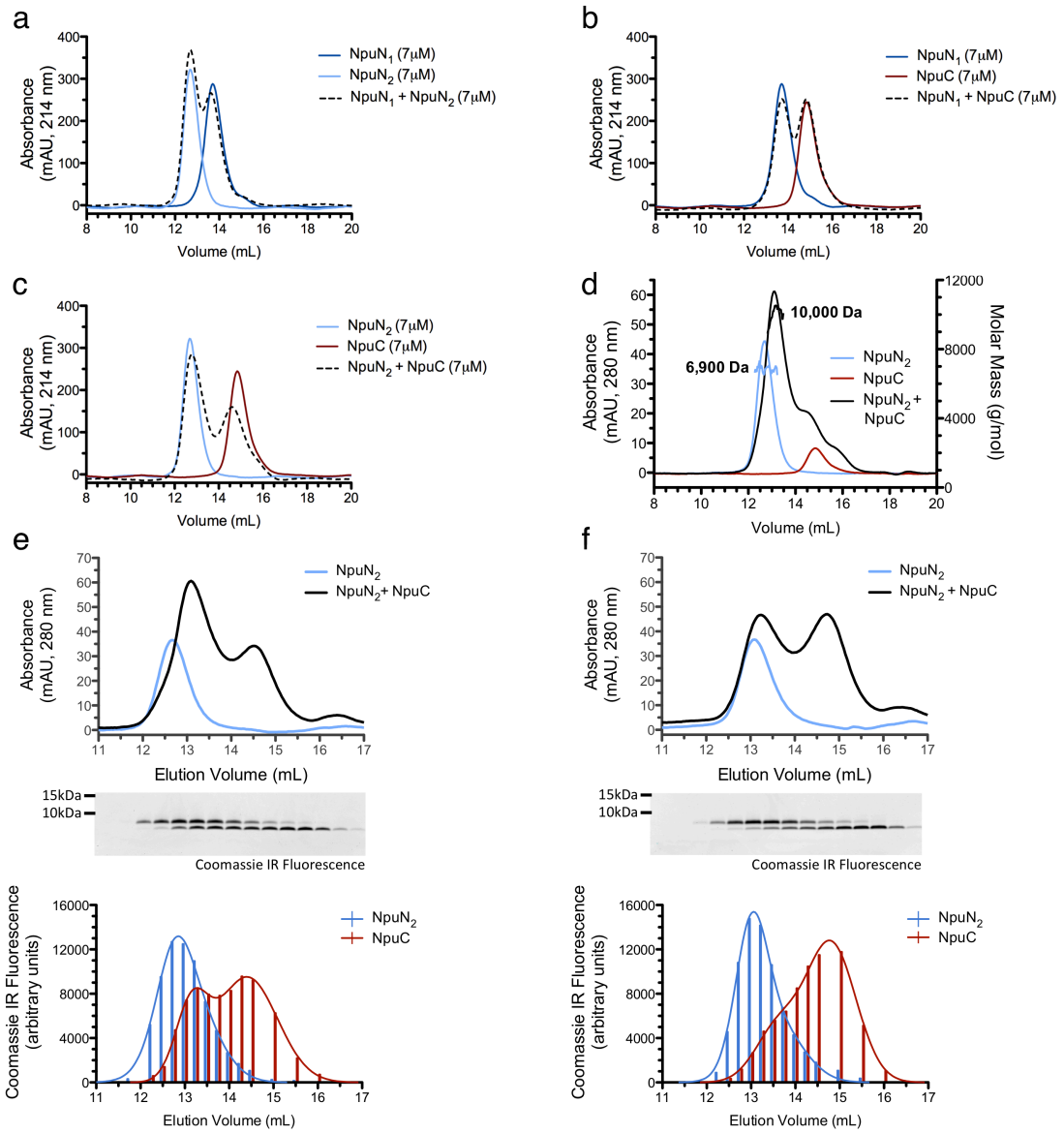


Figure 5.9. NpuN lobe interactions with NpuC. All SEC analysis was carried out in pH 6.5 buffer containing 25 mM sodium phosphates, 100 mM NaCl, and 1 mM DTT, except where noted. **a.** SEC chromatogram of a 7 μM $NpuN_1$ mixed with 7 μM $NpuN_2$. **b.** SEC chromatogram of a 7 μM $NpuN_1$ mixed with 7 μM NpuC. **c.** SEC chromatogram of a 7 μM $NpuN_2$ mixed with 7 μM NpuC. **d.** SEC-MALS of isolated $NpuN_2$ (light blue, 62.5 μM, expected MW = 6098 Da), NpuC (red, 7 μM), and an equimolar mixture of $NpuN_2$ and NpuC (black, 50 μM, expected MW = 10541 Da). Analysis of a 50 μM $NpuN_2$ mixed with 60 μM NpuC in **e.** 100 mM NaCl and **f.** 500 mM NaCl. Top panels show SEC chromatograms. Middle panels show SDS-PAGE analysis of select fractions collected throughout the elution of the major peaks in the black chromatogram. Bottom panels show histograms based on densitometric analysis of the coomassie-stained $NpuN_2$ and NpuC gel bands in each fraction. Histograms are fit to equations describing the sum of two Gaussian distributions.

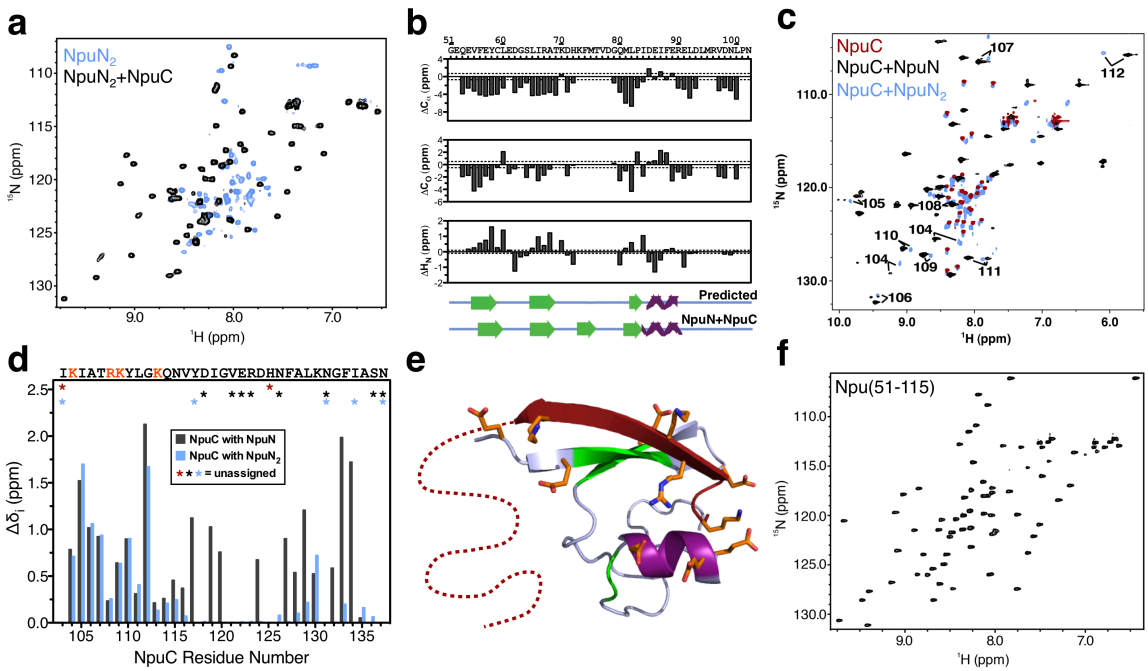


Figure 5.10. Structural characterization of the $NpuN_2$-NpuC complex. a. ^{1}H-^{15}N HSQC spectra of $NpuN_2$ alone (light blue, 250 μM) and with equimolar unlabeled NpuC (black, 250 μM) collected at 800 MHz. **b.** Chemical shift index values ($\Delta C/H = \delta C/H_{Observed} - \delta C/H_{RandomCoil}$) for Cα, CO, and HN atoms for $NpuN_2$ in complex with NpuC. The consensus secondary structure prediction is shown below compared with $NpuN_2$ in the context of the full Npu complex (as seen in PDB 2KEQ).[39] **c.** Overlay of ^{1}H-^{15}N HSQC spectra for NpuC alone (red) and in complex with $NpuN_2$ (light blue) or NpuN (black) at 800 MHz. Resonances corresponding to NpuC residues 104-112 are labeled to highlight the proximity of NpuN-complex (black) and $NpuN_2$-complex resonances (light blue). Note that in complex with NpuN and with $NpuN_2$, two NH cross-peaks are visible for residue 104. **d.** Composite ^{1}H and ^{15}N backbone chemical shift perturbation values ($\Delta\delta_i$) for NpuC in complex with NpuN (black) or $NpuN_2$ (light blue) calculated relative to isolated NpuC. The NpuC sequence is given above, with important cationic residues in orange. Below the sequence, an asterisk marks unassigned residues for isolated NpuC (red), the complex with NpuN (black), or the complex with $NpuN_2$ (light blue). **e.** Rendering of the $NpuN_2$-NpuC interaction, as seen in the native Npu complex (PDB 2KEQ). $NpuN_2$ is light blue, NpuC is red, and charged residues involved in intermolecular electrostatic interactions are highlighted as orange sticks. Predicted β-sheet and α-helical regions of $NpuN_2$ are colored in green and purple, respectively. Residues 103-115 of NpuC are rendered as a ribbon, and the remainder of the sequence is rendered as a dotted line. **f.** ^{1}H-^{15}N HSQC spectrum of Npu(51-115) collected at 600 MHz. NMR data was acquired at 25 °C in a pH 6.5 buffer containing 25 mM phosphates, 100 mM NaCl, and 1 mM DTT.

Since NpuC preferentially binds $NpuN_2$ over $NpuN_1$, we postulate that the $NpuN_2$-NpuC complex represents an intermediate in the coupled binding and folding pathway for DnaE inteins. Thus, we sought to characterize the structure of this complex. NMR spectroscopy of isolated $NpuN_2$ in the absence and presence of NpuC clearly indicated a global structural transition from a highly flexible state with no secondary structure to a collapsed state (Figure 5.10a). Analysis of backbone ^{1}H and ^{13}C chemical shifts for $NpuN_2$ in complex with NpuC indicated α-helical and β-sheet content in the same regions as in the native Npu complex, particularly for residues at the interface with NpuC (Figure 5.10b).[157,158] Unlike the global conformational change seen for $NpuN_2$ upon binding NpuC, only a fraction of the NpuC backbone resonances showed chemical shift perturbations upon binding $NpuN_2$ (Figure 5.10c). These new chemical shifts were similar to those found for NpuC in the full-length complex, suggesting native-like structure for these residues (Figure 5.10d). The perturbed residues primarily correspond to a contiguous stretch of nine amino acids at the N-terminus of NpuC (104-112), which make both ionic and β-strand pairing interactions with $NpuN_2$ in the full-length complex (Figure 5.10e). Indeed, the ^{1}H-^{15}N HSQC spectrum of a fusion between $NpuN_2$ and the first 13 residues of the C-intein, Npu(51-115), described a well-folded protein, indicating that this short stretch of NpuC was sufficient to collapse $NpuN_2$ (Figure 5.10f).

5.4. Split intein assembly is multi-phasic and electrostatically driven

The presence of a well-defined binding intermediate such as the $NpuN_2$-NpuC complex indicates that split intein assembly is multi-phasic and that these phases may be discretely observable. The fluorescence of the sole tryptophan in the Npu complex (W47,

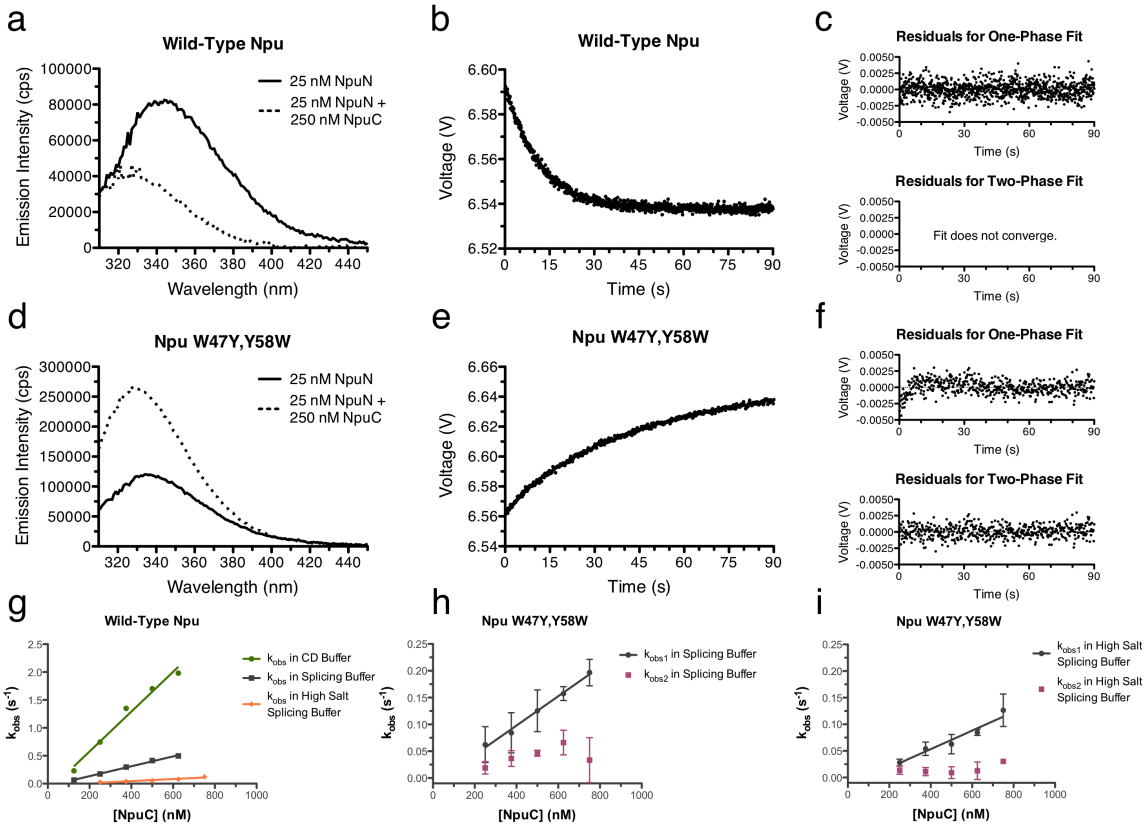


Figure 5.11. Intrinsic fluorescence binding measurements. a. Tryptophan fluorescence of wild-type NpuN in the absence and presence of NpuC (λ_{ex} = 290 nm). **b.** Tryptophan fluorescence of wild-type Npu upon stopped-flow mixing of fragments (λ_{ex} = 290 nm, λ_{em} = >320 nm). **c.** Residuals from fits to a one-phase (top) and two-phase (bottom) binding model for wild-type Npu. **d.** Tryptophan fluorescence of NpuN with the W47Y and Y58W mutations in the absence and presence of NpuC (λ_{ex} = 290 nm). **e.** Tryptophan fluorescence of Npu W47Y,Y58W upon stopped-flow mixing of fragments (λ_{ex} = 290 nm, λ_{em} = >320 nm). **f.** Residuals from fits to a one-phase (top) and two-phase (bottom) binding model for Npu W47Y,Y58W. **g.** Concentration-dependence of observed rates of wild-type Npu binding under pseudo-first order conditions with different buffers. **h.** Concentration-dependence of the observed rates of Npu W47Y,Y58W binding under pseudo-first order conditions with low salt. **i.** Concentration-dependence of the observed rates of Npu W47Y,Y58W binding under pseudo-first order conditions with high salt. For all stopped-flow measurements, [NpuN] = 25 nM and NpuC was mixed in 5- to 30-fold excess.

Figure 5.1c) is quenched upon Npu assembly and can be used to monitor binding (Figure 5.11a). Using this probe, we measured an equilibrium dissociation constant (K_D) for Npu of 1.2 ± 0.8 nM (Table 5.1). In kinetic measurements using stopped-flow fluorescence,

only a single binding phase was observable with a k_{on} of (8.9 ± 0.3) x 10^5 $M^{-1}s^{-1}$ and an off-rate too slow to extract from our kinetic data (Figures 5.11b,c,g and Table 5.2). W47 lies in $NpuN_1$, which our data indicate is compact throughout split intein assembly. Thus, we engineered a new tryptophan in the more dynamic $NpuN_2$ (Y58W, Figure 5.1c) and mutated W47 to tyrosine to silence $NpuN_1$ fluorescence. These mutations modestly affected the binding equilibrium (K_D = 3.4 ± 1.0 nM, Table 5.1) and did not significantly affect splicing activity (Table 5.3). Unlike W47 fluorescence, W58 fluorescence increased dramatically upon fragment association, consistent with $NpuN_2$ undergoing a folding transition upon binding NpuC (Figure 5.11d). Stopped-flow fluorescence of this mutant divulged two binding phases (Figures 5.11e,f,h and Table 5.2), an initial concentration-dependent fast phase with a k_{on} of (2.7 ± 0.4) x 10^5 $M^{-1}s^{-1}$ followed by a slower concentration-independent phase (k_{obs} = 0.04 ± 0.02 s^{-1}) that was only marginally faster than the overall protein splicing rate (k_{splice} = ~0.01 s^{-1}). Notably, for both the wild-type and mutant inteins, the K_D increased and k_{on} decreased in higher ionic strength buffers (Tables 5.1 and 5.2 and Figure 5.11g-i). These data are consistent with a rapid, electrostatically-driven encounter between two fragments followed by a collapse of the structure.

Table 5.1. Equilibrium dissociation constants from steady-state titrations.

Npu	Exteins	Buffer	K_D (nM) [a]
WT	Ub-SUMO	100 mM P_i, 150 mM NaCl	2.9 ± 1.8 [b]
WT	Ub-SUMO	100 mM P_i, 500 mM NaCl	13.3 ± 2.6
WT	AEY-CFN	100 mM P_i, 150 mM NaCl	< 1 [c]
WT	AEY-CFN	100 mM P_i, 500 mM NaCl	1.2 ± 0.8
W47Y,Y58W	AEY-CFN	100 mM P_i, 500 mM NaCl	3.4 ± 1.0

[a] Errors indicate the standard error in the best-fit K_D for a global fit of 3-4 datasets to the quadratic binding equation.
[b] This value was previously determined (described in Chapter 4).
[c] With small exteins in the 150 mM NaCl buffer, the fragment affinity was too tight to accurately extract a K_D from the fluorescence titration curve.

Table 5.2. Rate constants extracted from stopped-flow measurements.[a]

Npu	Buffer Name	Buffer	k_{on} ($x10^5$ M^{-1} s^{-1}) [b]	k_{obs2} ($x10^{-2}$ s^{-1}) [c]
WT	CD Buffer	100 mM P_i, 50 mM NaF	35.6 ± 3.0	-
WT	Splicing Buffer	100 mM P_i, 150 mM NaCl	8.9 ± 0.3	-
WT	High Salt Buffer	100 mM P_i, 500 mM NaCl	1.9 ± 0.3	-
W47Y,Y58W	Splicing Buffer	100 mM P_i, 150 mM NaCl	2.7 ± 0.4	4.0 ± 2.4
W47Y,Y58W	High Salt Buffer	100 mM P_i, 500 mM NaCl	1.7 ± 0.2	1.4 ± 1.1

[a] Errors indicate the standard error in the best-fit rate constant for a global fit of 1-4 datasets.
[b] k_{on} was determined as the slope of the concentration-dependent k_{obs} line (Figures 5.11g-i).
[c] For the biphasic reactions, the reported k_{obs2} is the average value for all observed k_{obs2} values at all NpuC concentrations (i.e. average of all magenta data points in Figures 5.11h,i).

Table 5.3. Rate constants for splicing of Npu tryptophan mutants.[a]

Npu	k_1 (s^{-1})	k_2 (s^{-1})	k_3 (s^{-1})	k_{splice} (s^{-1})
WT	(5.21 ± 0.28) x 10^{-2}	(1.77 ± 0.38) x 10^{-2}	(3.15 ± 0.04) x 10^{-2}	(1.36 ± 0.02) x 10^{-2}
W47Y	(5.27 ± 0.21) x 10^{-2}	(1.71 ± 0.30) x 10^{-2}	(3.78 ± 0.10) x 10^{-2}	(1.56 ± 0.03) x 10^{-2}
Y58W	(4.24 ± 0.25) x 10^{-2}	(1.30 ± 0.19) x 10^{-2}	(3.60 ± 0.07) x 10^{-2}	(1.36 ± 0.04) x 10^{-2}
W47Y,Y58W	(3.72 ± 0.24) x 10^{-2}	(1.24 ± 0.16) x 10^{-2}	(4.21 ± 0.01) x 10^{-2}	(1.42 ± 0.01) x 10^{-2}

[a] Rate constants were determined using the assay and kinetic model described in Chapter 3.

5.5. Split intein fragments associate via a "capture and collapse" mechanism

Based on the foregoing biophysical experiments, we propose that Npu fragment assembly is a multi-step process in which NpuC is first engaged by $NpuN_2$, followed by $NpuN_1$ (Figure 5.12). The initial encounter between the fragments is dictated by electrostatic interactions between the disordered, highly cationic NpuC and the extended, highly anionic second lobe of NpuN. This encounter complex resolves to an intermediate that has secondary structure and intermolecular interactions resembling half of the native Npu complex. After this initial "capture", the flexible regions "collapse" into an ordered state that is further stabilized by hydrophobic interactions between NpuC and the well-ordered first lobe of NpuN. The resulting complex has intertwined fragments, and this topology is presumably only accessible due to the intrinsic disorder of the starting materials. This "capture and collapse" model is meant to provide a framework for

understanding intein fragment assembly, and we recognize that there could be additional folding steps within each phase that are not detectable using the biophysical approaches employed herein.

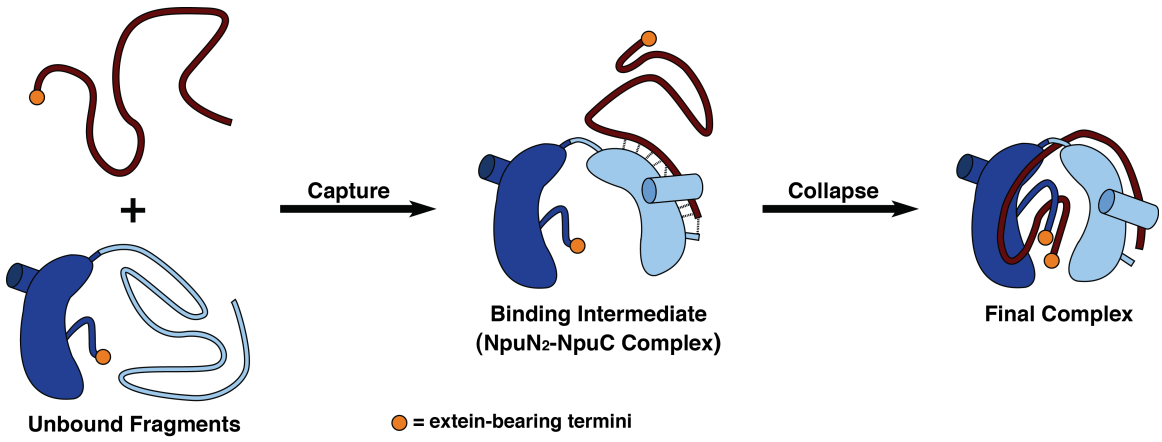


Figure 5.12. The "capture and collapse" mechanism of split intein assembly. An N-terminal segment of the disordered C-intein is captured by the extended second lobe of the N-intein with compaction of that lobe into a native-like structure. This intermediate then collapses as the remainder of the C-intein docks into the pre-organized first lobe of the N-intein.

5.6. Native topology is important for split intein stability and function

NpuC intercalates between the NpuN lobes, allowing it to engage $NpuN_2$ through electrostatic interactions and $NpuN_1$ through hydrophobic interactions (Figure 5.6). These segregated intermolecular interactions and the resulting interweaving of the chains should not only drive assembly, as described above, but also prevent the bound fragments from rapidly dissociating, thereby precluding splicing. To test this notion, we engineered two Npu variants with altered primary sequence topology. In the first system, we permuted the split intein sequence to yield two new proteins related by a pseudo-symmetry axis (Figures 5.1d and 5.13a). One protein, Npu(51-115) described above, represents the well-folded binding intermediate comprised of $NpuN_2$ and a segment of NpuC, where all crucial ionic interactions are pre-satisfied (Figure 5.10e,f). The second protein, a fusion

of $NpuN_1$ and the remainder of NpuC, pre-satisfies the hydrophobic interactions between these two protomers (Figure 5.6a,b). A complex of these two permuted fragments should associate slowly and dissociate rapidly, given the lack of key intermolecular interactions and entwined fragments. As expected, the permuted intein fragments could assemble into the correct intein fold, as evidenced by their capacity to carry out protein splicing (Figures 5.13b-e). However, the rate of splicing was extremely concentration dependent, indicating that association is rate-limiting. Furthermore, even at high concentrations, splicing was dramatically slower than wild-type activity (days instead of minutes), which we attribute to a short lifetime of the active complex due to a fast off-rate.

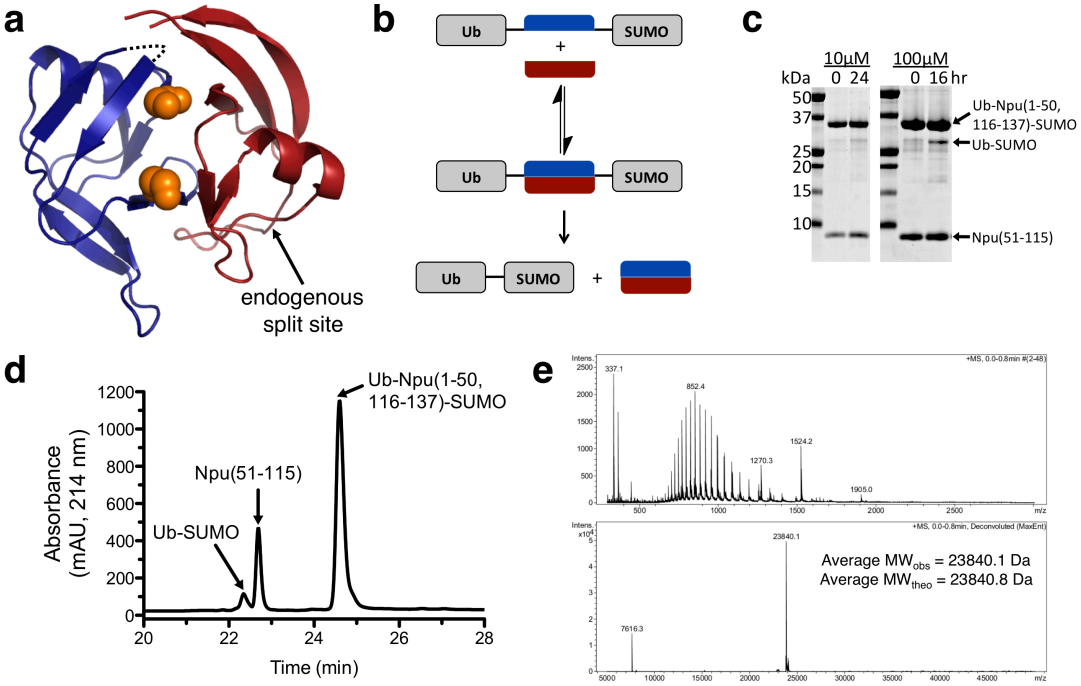


Figure 5.13. Activity of the permuted Npu complex. a. Structural representation of the two intein regions in the permuted Npu constructs (based on PDB 2KEQ). The dashed line represents a diglycine linker introduced between residues 1-50 and 116-137. **b.** Scheme showing protein splicing with the permuted Npu. Ub (ubiquitin) and SUMO are model N- and C-exteins **c.** Proteins splicing of permuted Npu at 10 μM and 100 μM fragment concentrations. Reaction mixtures where analyzed at indicated time-points by SDS-PAGE with coomassie staining. **d.** RP-HPLC analysis of the endpoint of the 100 μM reaction shown in panel c. Peaks were identified by mass spectrometry. **e.** Mass spectrum of the Ub-SUMO product.

In the second system, we utilized the isolated NpuN lobes in conjunction with NpuC to generate a three-piece intein. While these constructs must contend with the entropic cost of forming a ternary complex, the key intermolecular interactions between each lobe and NpuC are still present. The three-piece complex was stable at low-micromolar concentrations (Figure 5.14a) where the binary interaction between $NpuN_2$ and NpuC was barely visible (Figure 5.9c). Since NpuC cannot independently interact with $NpuN_1$ (Figure 5.9b), this result shows that the lobes bind NpuC cooperatively. Notably, the resulting ternary complex could carry out protein splicing with substantially higher efficiency than the permuted complex (Figures 5.14b-f), demonstrating the kinetic stability of the active complex.

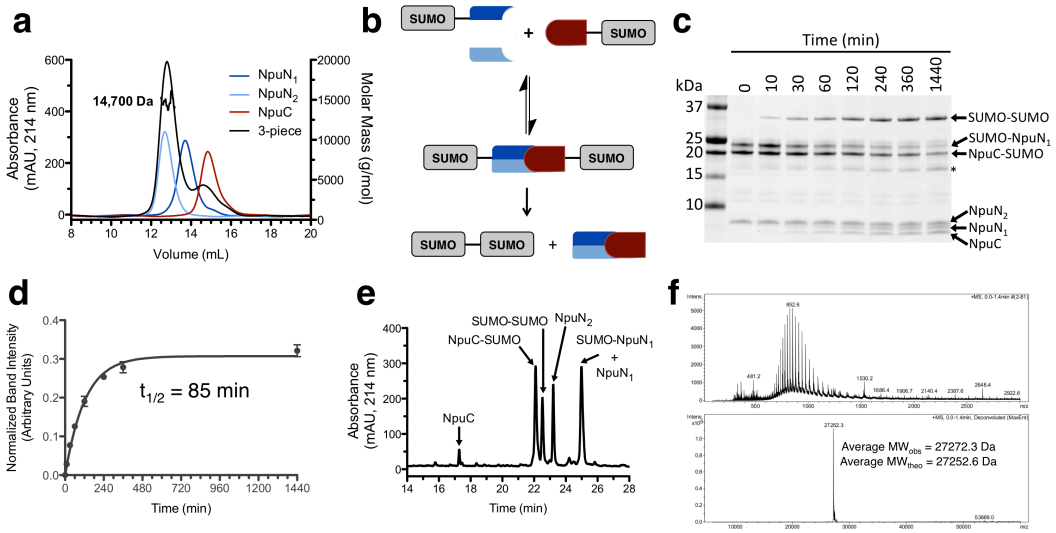


Figure 5.14. Stability and activity of a three-piece intein. a. SEC-MALS of the three-piece Npu complex (black, 7 µM of each fragment, expected MW = 16649 Da) compared with the individual components (each at 7 µM). **b.** Scheme showing protein splicing with the three-piece Npu. **c.** Splicing activity of the three-piece Npu at 5 µM fragment concentrations. Reaction mixtures where analyzed at indicated time-points by SDS-PAGE with coomassie staining. The asterisk indicates a side product, most likely SUMO due to premature N-extein cleavage. **d.** Reaction progress curve for the formation of the SUMO-SUMO product. The reaction was analyzed in triplicate, and error bars represent the standard deviation. The half-life of 85 minutes is based on a fit to a first-order rate equation. **e.** RP-HPLC analysis of an intermediate time point in the reaction where all starting materials and products are visible. Peaks were identified by mass spectrometry. **f.** Mass spectrum of the SUMO-SUMO product, confirming its identity.

5.7. Summary and conclusions

Inteins have a unique fold in which the polypeptide chain traverses back and forth between two symmetry-related halves (Figure 5.1c). Given this, it is remarkable that some inteins are naturally split, as this topology precludes a binding mechanism involving natively structured fragments. In this report, we describe the structure of split intein fragments and determine their mechanism of association. Using a variety of biophysical techniques coupled with protein engineering and protein chemistry, we show that the N-intein fragment has a unique bipartite structure comprised of one at least partially folded lobe tethered to an intrinsically disordered region of equal length. This extended region of the N-intein captures the completely disordered C-intein through electrostatic interactions, leading to a native-like intermediate. This intermediate then further collapses onto the more structured N-intein lobe, resulting in a functional intein domain with entangled fragments.

In order to function, split intein fragments must not only find one another efficiently, but also bind to form a stable complex that can persist throughout the duration of the complex multi-step splicing reaction, which can last from seconds to hours (as seen in Chapter 2).[123] Our studies indicate that split DnaE inteins have evolved several unique structural features that fulfill these requirements: (1) The initial encounter complex is dictated by two disordered protein regions, which should provide larger capture radii for an enhanced collision probability.[159] (2) This enhanced on-rate is further reinforced by electrostatic steering of these fragments.[129] (3) The slow off-rate is presumably governed by the retention of these ionic interactions in the final complex (discussed in Chapter 4),[98] as well as significant hydrophobic interactions between NpuC and $NpuN_1$ (Figure 5.6a,b).

(4) The disordered regions of the fragments provide access to an interwoven topology, which further reduces the off-rate (Figure 5.1c).

The significance of this interweaving for intein function was demonstrated using two engineered systems: one with a permuted primary sequence and another involving a three-piece intein. It is noteworthy that both of these systems could, in principle, be used to design novel intein-based technologies. The permuted system, with poor binding, could be enhanced by fusion to an inducible dimerization system, thus making a conditionally splicing Npu. While such a system has never been developed with a highly efficient intein such as Npu, conditional protein splicing modules show promise as tools to control protein function *in vivo* in response to external stimuli (discussed in section 1.4.5).[82] The three-piece system is particularly intriguing, as it is implicitly conditional: the central piece, $NpuN_2$, is effectively a trigger for protein oligomerization and splicing. Further design and optimization of both the permuted and three-piece Npu inteins should lead to the development of useful tools for biochemistry and cell biology.

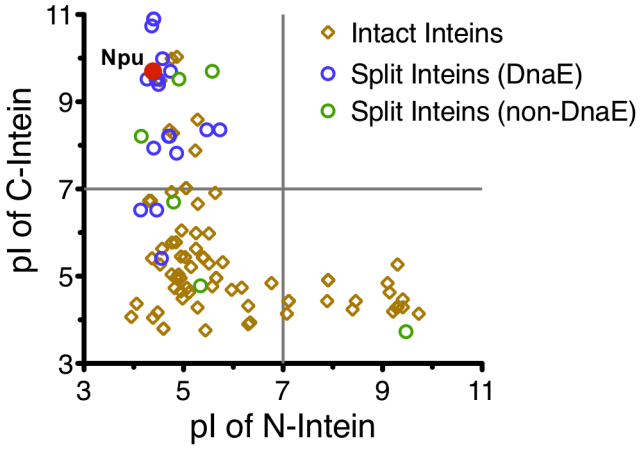


Figure 5.15. Charge segregation in diverse split inteins. Plot of theoretical isoelectric points for N- and C-terminal fragments of naturally split inteins (circles) and the corresponding homologous regions from naturally contiguous inteins bearing no endonuclease domain (diamonds). The dataset used was adapted from our previous work (Chapter 4),[98] and several new non-DnaE split inteins were added.[17]

Split inteins likely arose from their contiguous predecessors due to DNA inversions or aberrant gene insertions and excisions.[17,19] These genomic rearrangements independently occurred several times, as evidenced by the diversity of split inteins, their host proteins, and their host organisms.[17,19,20] Interestingly, even outside of the DnaE family described in this study, many split inteins are cut at the same point on the intein fold and exhibit dramatic charge segregation between fragments (Figure 5.15). Furthermore, homology modeling of these divergent split inteins predicts that many have intermolecular electrostatic interactions at the ridge of their horseshoe-like structure, as seen in the DnaE inteins.[17] Thus, the efficient folding mechanism for Npu is most likely a general solution at the protein level to the rare but repeated fracturing of intein genes at the DNA level. Given that split inteins assemble essential proteins (e.g. DNA polymerase alpha subunit, DNA helicase, and ribonucleotide reductase) it is not surprising that this solution entails a number of traits that ensure highly efficient binding. Ultimately, the molecular details of this mechanism, laid out in this study, not only provide insight into the evolution of efficient biomolecular recognition, but they will also lay the groundwork for future engineering efforts on this unusual and useful class of proteins.

Chapter 6: Discussion and outlook

The work presented in this thesis describes the detailed biochemical and structural characterization of naturally split inteins and explores the utility of these molecules as efficient tools for protein engineering. Prior to the discovery of the fast-splicing Npu intein,[89] it was broadly believed that there is a slow speed limit for protein splicing dictated by the chemistry involved, especially peptide bond cleavage and nucleophilic attack by an asparagine side chain. The (apparently) unique kinetic properties of Npu have been revolutionizing intein-based technologies, and they also spurred the basic research described herein. In this chapter, the immediate practical implications of this basic research are discussed. Then, persistent shortcomings of split intein applications are addressed, focusing on ways that the biochemical and biophysical findings presented in the previous chapters may be used to overcome these caveats. Finally, some open questions in intein research and possible future experiments are presented.

6.1. Practical implications of the discovery and design of new inteins

As indicated throughout this thesis, naturally split inteins can potentially enhance intein-based protein engineering techniques. Their tight binding affords the capacity for protein ligation under native conditions at low concentrations. The fact that this assembly relies on a complex protein-protein interaction means that it is highly specific, allowing for protein semi-synthesis to be done in living cells and organisms. Furthermore, whereas efficient protein and peptide cyclization using contiguous inteins in Expressed Protein Ligation depends on the proximity of termini, cyclization by protein *trans*-splicing is robust and can provide access to large libraries of genetically encoded cyclic peptide

libraries. All of these enhancements to intein-based technologies were originally made using the slow Ssp split intein, and so it is not surprising that second generation versions of these techniques are now employing the fast but homologous Npu intein.[86,93,125]

Our work, described in Chapter 2, indicates that the "ultrafast" Npu intein can also be used to improve the flagship intein-based technology, Expressed Protein Ligation.[56] The fact that this intein has a highly activated N-terminal splice junction means that it provides quicker access than the commercially available GyrA intein to proteins with a C-terminal α-thioester moiety, which can be used for downstream ligation chemistry (Figures 2.10 and 2.11).[123] This hyperactivation comes at a cost, as fusions to contiguous forms of Npu are more susceptible to premature cleavage as well (Figure 2.12). The streamlined EPL technology (sEPL) overcomes this caveat by using split versions of Npu, which have a latent thioester that is only activated upon fragment binding. As binding can also be used for affinity enrichment, sEPL is a quick method to obtain pure α-thioesters of recombinant proteins (Figure 2.14).[124]

Our initial results to enhance EPL with Npu indicate that multiple aspects of this protein (tight binding and hyperactivated N-terminus) are extremely useful. What, then, is the value in discovering other fast-splicing inteins? From a practical standpoint, we found that while Npu has some remarkable properties, fusions to the NpuN fragment do not always express well in bacterial or mammalian cells (Figures 2.8 and 2.17). Here, our newly-discovered library of fast-spicing N- and C-inteins come in handy, as the fragments can be screened for suitable expression levels when fused a protein of interest. Since these fragments cross-react while retaining their fast splicing (and apparently their tight binding), any cognate or non-cognate pair can be chosen for a protein engineering

project based solely on expression levels (Figure 2.9). Indeed, for one major application of sEPL, the generation of site-specificially modified monoclonal antibodies (mAbs), we find that the AvaN fragment and NpuC fragment are a well-suited combination (Figure 2.17), whereas NpuN fusions to mAbs express poorly, and AvaC cannot be readily generated in high yields and immobilized onto a solid support.*

The discovery of multiple fast-splicing naturally split inteins opens the door to another interesting technological prospect: carrying out multiple ligations in one-pot. This type of reaction scheme could be useful for *in vivo* protein semi-synthesis experiments where labeling of multiple proteins with different tags (e.g. different colored fluorescent dyes) is desired or for segmental isotopic labeling of central regions in multi-domain proteins. Unfortunately, the DnaE inteins described in Chapter 2 are not suitable for this application, as the orthologous fragments cross-react (Figure 2.9). By contrast, our engineered charge-swapped Npu intein, described in Chapter 4, can be used in conjunction with wild-type Npu for parallel selective *trans*-splicing reactions (Figure 4.5), although the splicing kinetics of the mutant intein are more concentration-dependent than the native counterpart (Figure 4.4). Furthermore, the reaction specificity for the wild-type and charge-swapped Npu is under kinetic control and is not absolute,[98] which may even limit the scope of this system for *in vitro* applications such as segmental isotopic labeling.

Recently, several non-DnaE split inteins were identified in a metagenomics study[17] and their splicing kinetics were found to be extremely rapid, on the order of tens-of-seconds to minutes.[160] Furthermore, as several of these are non-allelic and thus have

* Ongoing work and improvement of the sEPL technology, especially for the generation of antibody-drug conjugates, is being carried out by Miquel Vila-Perelló.

low sequence homology, they do not cross-react. While it remains to be seen if these new inteins bind as tightly as the DnaE inteins, a requirement for their use *in vivo*, our biophysical studies on Npu, coupled with sequence analyses, suggest that this should be a ubiquitous property of all naturally split inteins (Chapter 5).

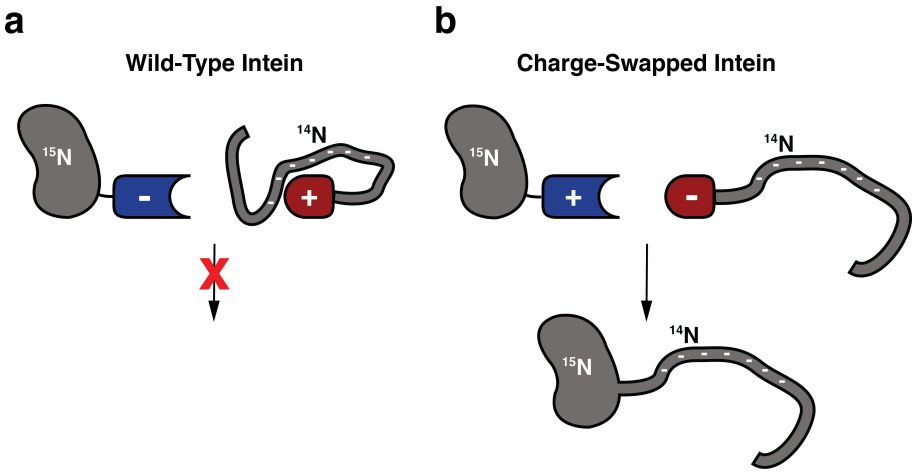


Figure 6.1. Segmental labeling with the charge-swapped intein. If electrostatic interactions between an extein and the intein fragment it is attached to inhibit protein splicing, using the charge-swapped intein designed in Chapter 4 can circumvent this problem.

The charge-swapped Npu intein may require further optimization before its use *in vivo*, but it does have one immediate practical application. As split inteins are increasingly being used for *in vitro* protein semi-synthesis, especially for segmental isotopic labeling, there is a growing body of empirical evidence indicating that split inteins are extremely inefficient when fused to certain protein domains or fragments (different from the local sequence specificity described in Chapter 3). This is presumably due to some unfavorable non-specific interaction between the intein fragment and its cargo. As split intein protomers are highly charged, these inhibitory interactions may be electrostatic (Figure 6.1a). In fact, segmental labeling is a potentially useful approach to study proteins that have folded regions attached to intrinsically disordered regions, the

latter of which typically have a high charge density of their own.[3] In cases like these, if *trans*-splicing efficiency of native DnaE inteins is hampered, sample preparation may be simplified by the use of the engineered Npu with inverted charges (Figure 6.1b).

6.2. Persistent challenges with split intein technologies

The two major caveats that have inhibited the use of inteins for protein engineering are their slow reactivity and their extein sequence-dependent splicing activity. With the discovery and characterization of roughly 15 split inteins that can splice in minutes or less, the first caveat is effectively solved (Chapter 2).[123,160] However, this is only true when these inteins are tasked with carrying out protein splicing or cleavage reactions at sequences resembling their endogenous substrate. Out of this optimal context, even these "ultrafast" split inteins can have reaction kinetics and yields as poor as the previously studied inefficient inteins (Chapter 3). Extein dependence is a pervasive issue that affects almost all technologies that require protein *trans*-splicing. Other more technology-specific issues include the synthetic inaccessibility of split intein fragments for protein semi-synthesis with non-proteinogenic cargo and the fact that there is no way to trigger the assembly of highly efficient naturally split inteins for conditional protein splicing. Given the information presented in this thesis, some of these *trans*-splicing drawbacks may now be addressed.

6.2.1. Engineering a "promiscuous" Npu split intein

Npu and other DnaE inteins show a strong dependence on the identity of local C-extein residues, with the most efficient splicing occurring in the presence of sequences

resembling their endogenous splice junction (C-F-N at the +1 to +3 positions).[90,91,128] In particular, as discussed in Chapter 3, Npu requires a bulky residue at the +2 position, and removal of this steric bulk results in roughly a 100-fold decrease in the rate-limiting branched intermediate resolution step (Figure 3.7).[151] We found that this is due to the Phe_{+2} side chain plugging the intein active site, thus forcing a catalytic His_{125} side chain and the C-intein/C-extein junction in close proximity (Figures 3.9 and 3.10). Based on our structural studies, His_{125} sits on a flexible loop, and we reasoned that the effects of removing steric bulk at the +2 position could be compensated for by re-engineering this loop to constrain active-site motions from *within* the intein. In fact, there are several examples of point mutations on this loop either slightly alleviating C-extein-dependence (as shown by us in Chapter 3 for Npu and by others on different inteins)[92,95,151] or improving overall splicing activity (as demonstrated for Ssp in Chapter 2 and in previous work on an NpuN-SspC chimera).[92,123]

In these previous efforts mutations were identified either by rational design based on sequence alignments and previous reports or by directed evolution using error-prone PCR with a low mutation rate over the entire intein sequence. Since our structural studies indicate that the His_{125} loop is a hot-spot for controlling branched intermediate resolution, we recently designed an experiment to site-specifically mutagenize this loop and select for relaxed C-extein specificity.* In this experiment, we first designed an oligo-nucleotide DNA primer spanning this loop with a degenerate mixture of bases at the codons for positions 122-124. Using standard PCR techniques, we generated a library of Npu genes in which all possible amino acid sequences were sampled within this loop. Then, using

* This work was initiated in collaboration with Adam Stevens, and further efforts to engineer and characterize "traceless" inteins will be continued by him.

the intein activity-coupled kanamycin resistance approach described in Chapters 2-4,[92] we selected bacterial clones that could survive in the presence of high kanamycin concentrations when the local C-extein sequence was "CGN" rather than "CFN" (Figure 6.2a). The clones were sequenced, and several Npu mutants were analyzed *in vitro*.

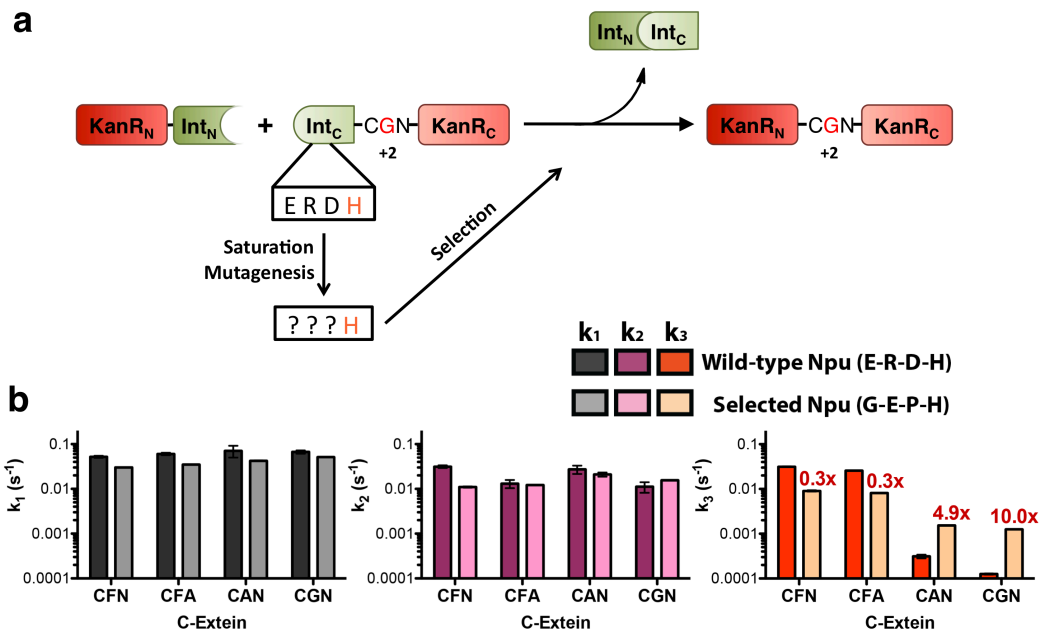


Figure 6.2. Saturation mutagenesis of the His_{125} loop to find a traceless intein. a. Scheme depicting the kanamycin selection system to identify loop mutants that tolerate a +2 glycine. **b.** Rate constants for *in vitro* splicing with the wild-type intein and one containing the E122G, R123E, and D124P mutations. k_1 is the rate of branched intermediate formation, k_2 is the rate of the reverse reaction, and k_3 is the rate of branched intermediate resolution. The red numbers above the k_3 bars indicate the fold-difference between wild-type and the selected mutant.

The results of this preliminary saturation mutagenesis experiment are extremely promising. In one instance, we isolated an Npu mutant in which the native loop sequence "ERD" was mutated to "GEP" (Figure 6.2b). As predicted from our structure-activity studies in Chapter 3, these mutations had virtually no effect on the initial steps (k_1 and k_2) in the splicing reaction but dramatically affected branched intermediate resolution (k_3). This step was 10-fold faster with the mutant intein than wild-type in the presence of

Gly_{+2}, and the loop mutations even improved reactivity with an Ala_{+2} 5-fold (bringing both of these reactions into roughly a 10 minute time-frame). Importantly, the mutant intein was still tolerant of the bulky Phe_{+2} residue: there was a roughly 3-fold decrease in branch resolution kinetics, which still afforded protein *trans*-splicing in a few minutes. Furthermore, the mutations had no effect on the inherent promiscuity of Npu towards the +3 position (Figure 3.7), suggesting that this position is sufficiently far from the active site to matter in any context.

Future experiments will focus on whether or not these loop mutations introduce some new specificity towards N-extein sequences and also on characterizing the structural mechanism by which the mutations evade C-extein dependence. In particular, it will be interesting to see if these mutations recapitulate the constraints on active-site motions described for Phe_{+2} in Chapter 3 (Figures 3.9 and 3.10).[151] Regardless of the precise mechanism, this engineered intein will unequivocally enhance the use of *trans*-splicing for *in vitro* and *in vivo* protein semi-synthesis and peptide cyclization, among other techniques.

6.2.2. Making synthetically accessible, rapid-splicing split intein fragments

As noted in Chapter 1, split inteins have been used for both *in vitro* and *in vivo* protein semi-synthesis endeavors. Given their tight binding, these fragments can be used to ligate proteins and peptides with relative ease. When the desired product has non-proteinogenic moieties or post-translational modifications, however, one of the fragments must be made synthetically.[58] Here, protein semi-synthesis by *trans*-splicing is more technically challenging, as natural N-inteins have more than 100 amino acids, making

them synthetically inaccessible, and C-inteins are roughly 35 amino acids. While the latter are synthetically accessible, the addition of any cargo puts these molecules nearly out of reach. To address these issues, several groups have generated split inteins with non-canonical split sites based on naturally split inteins or artificially split contiguous inteins with the goal of moving the split site closer to the N- or C-terminal splice junction.[39,59-61,161] While these efforts have yielded synthetically accessible pieces, the engineered split intein fragments have slow association kinetics and weaker binding affinities, which in turn reduce splicing efficiency.[62]

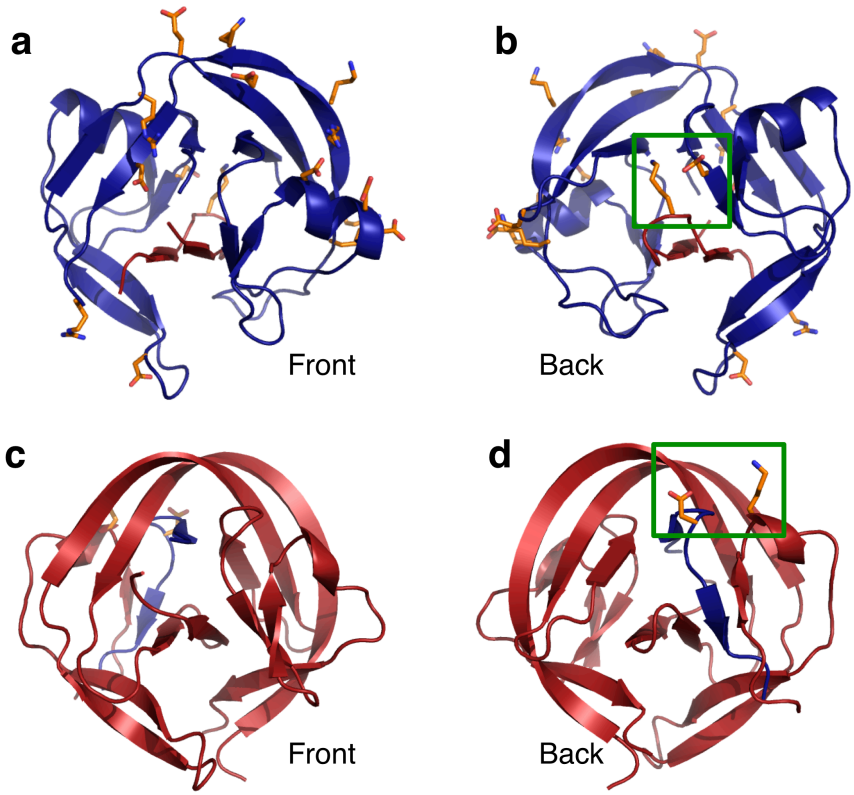


Figure 6.3. Electrostatic interactions in non-canonically split inteins. a. and **b.** Renderings of NpuDnaE split 15 residues from the C-terminus (PDB 2KEQ). **c.** and **d.** Renderings of SspDnaB split 11 residues from the N-terminus (PDB 1MI8). N- inteins are blue and C-inteins are red. The residues known to participate in intermolecular salt bridges in native Npu are shown as sticks in panels a and b. In panels b and d, the only salt bridges retained in the non-canonically split inteins are highlighted in green boxes.

Our characterization of Npu fragment assembly in Chapters 4 and 5 provides insights into why these approaches fall short of the desired outcome. When truncating the C-intein and elongating the N-intein of naturally split inteins as has been done with Npu,[60] several crucial *inter*molecular electrostatic interactions become pre-satisfied *intra*molecular interactions (Figure 6.3a,b). These interactions, at the interface between NpuC and the second lobe of NpuN ($NpuN_2$), are involved in the initial recruitment step for binding as well the stability of the final complex (Table 4.2 and Figure 5.10). By contrast, moving the split site close to the N-terminal splice junction, which has only been successfully implemented with an artificially split DnaB intein,[61] results in fragments with few stabilizing intermolecular interactions beyond shape complementarity, hydrogen bonding between two short β-strands, and one salt bridge (Figure 6.3c,d). Given the importance of electrostatic interactions in the assembly of naturally split inteins, an interesting prospect to improving non-canonically split inteins through rational design would be to simply introduce more intermolecular ion pairs at the interface.

Realistically, it may be impossible to recapitulate the low-nanomolar affinities observed for Npu and Ssp without the topological entangling between fragments to provide a large buried surface area and a slow off-rate. An alternative approach would be to improve the synthetic methods for producing intein fragments. One approach, already explored for mechanistic studies in Chapter 3, is to create C-intein conjugates to C-exteins semi-synthetically, thereby avoiding the need to chemically synthesize the 35 amino acid C-intein.[151] In this strategy, the full NpuC sequence was expressed in *E. coli* as a fusion to the contiguous GyrA intein then thiolyzed off to yield an NpuC-MES

thioester at the C-terminal Block G asparagine residue. Asparagine thioesters are chemically unstable, and so the NpuC-MES peptide was reacted with an N-terminal cysteine-containing peptide *in situ* to yield a ligated product (Figure 6.4a).

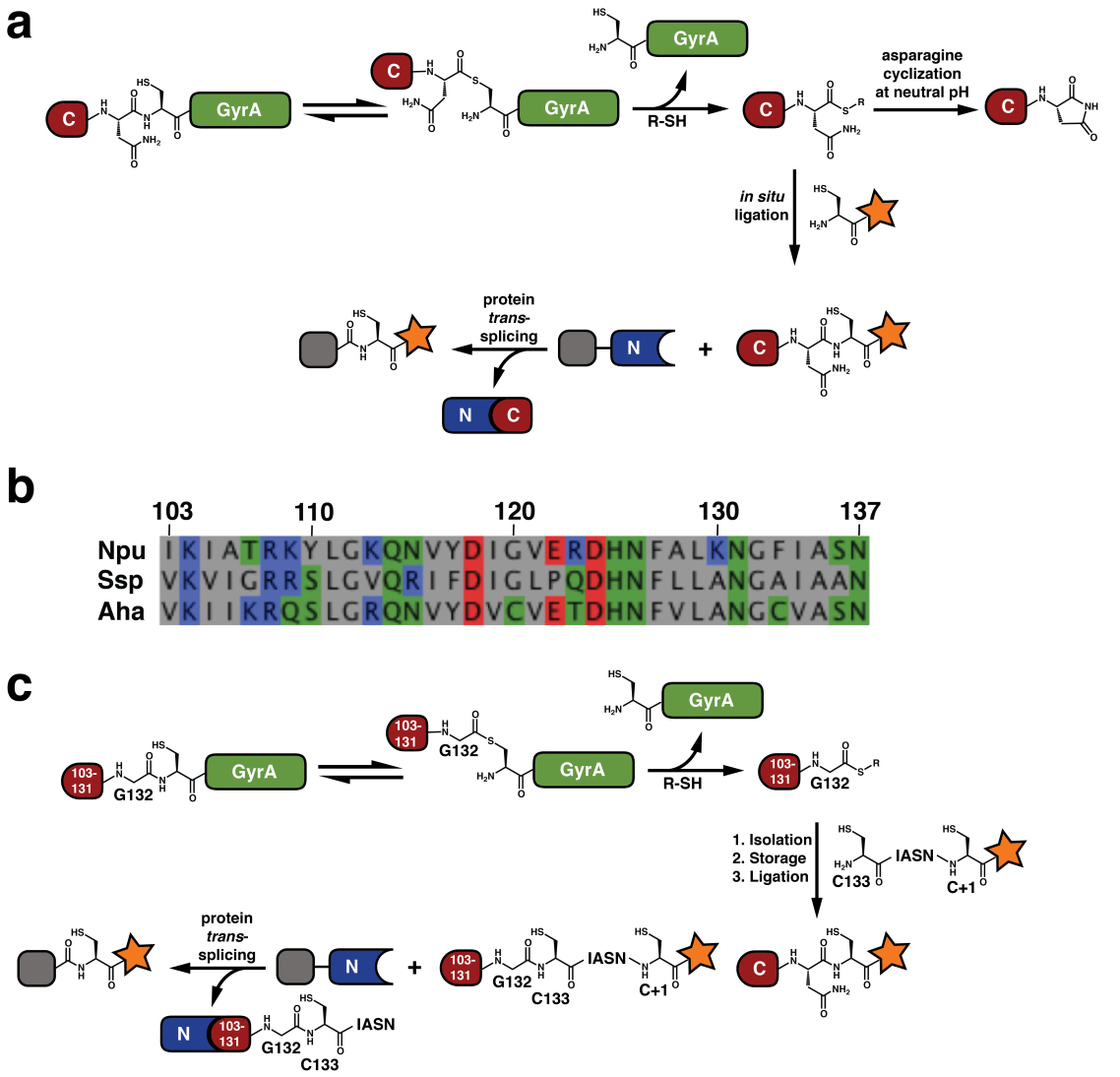


Figure 6.4. Semi-synthetic approaches to making C-inteins with synthetic cargo. a. The synthetic scheme utilized in Chapter 3 to make C-inteins with a variety of C-exteins. This approach requires a one-pot thiolysis and ligation, as the asparagine thioester is not readily isolable. **b.** Sequence alignment of the Npu, Ssp, and Aha C-inteins showing two cysteines in AhaC at positions 120 and 133. **c.** Semi-synthetic scheme to make a C-intein with a synthetic cargo by utilizing NpuC with an F133C mutation.

To circumvent the need for an asparagine thioester, one approach would be to introduce a cysteine within NpuC, close to the C-terminus. A close look at the sequence

alignment of DnaE inteins indicates that the Npu ortholog from *Aphanothese halophytica* (Aha) has two cysteines within its C-fragment (Figure 6.4b). One of these, discussed in Chapter 2, causes aberrant N-extein cleavage but can be replaced with a glycine to restore high activity *in vivo* (Figure 2.6).[123] The second is only five residues from the C-terminus (position 133). Thus, one could generate either an NpuC or AhaC_{C120G} intein missing the last five residues (103-132) as a GyrA fusion and isolate the thioester at glycine 132. This truncated Int$C_{103-132}$-MES could be mass-produced, stored, and ultimately reacted with any synthetic peptide that starts with a cysteine and four additional C-terminal IntC residues (133-137) followed by the crucial +1 cysteine and any desired synthetic cargo (Figure 6.4c).

6.2.3. Designing a highly efficient intien with controllable binding

Conditional protein splicing (CPS) is probably the most intriguing intein-based technology. It is premised on the notion that by controlling an intein's function using some external factor, one can trigger the function of the spliced product "on demand".[82] The most intuitive CPS frameworks control the hetero-dimerization of artificially split inteins, as in the absence of complex formation there is no basal splicing or side reactions. To date, the CPS systems that have been developed using this approach are either low-yielding or slow, as they rely on inefficient, poorly-behaved intein fragments.[81-83] Thus, it would be desirable to control the binding of a fast-splicing naturally split intein with well-behaved protein fragments.

Npu splicing is extremely fast and fragment binding is rapid, tight, and spontaneous. Remarkably, our charge-swapping mutagenesis experiments in Chapter 4

indicate that these properties are extricable.[98] In the $NpuN_{MUT}$ fragment, seven acidic residues were mutated to basic amino acids. This fragment could still bind the $NpuC_{WT}$ protomer (Figure 4.3), albeit with 70-fold weaker binding affinity than two wild-type fragments (~210 nM versus 3 nm). Given the seven mutations resulting in substantially weaker binding, at mid-nanomolar concentrations the $NpuN_{MUT}$-$NpuC_{WT}$ split intein took several hours to complete the splicing reaction, however at low-micromolar concentrations it could carry out protein *trans*-splicing in minutes (Figure 4.4). Thus, this designed fragment pair would be a good starting point for the development of a conditionally activated Npu intein. Splicing could be triggered using small molecule- or light-responsive dimerization domains.[81,83] Alternatively, such a system could be engineered to splice upon post-translational modification, for example to create a reporter for an enzyme of interest (Figure 6.5).

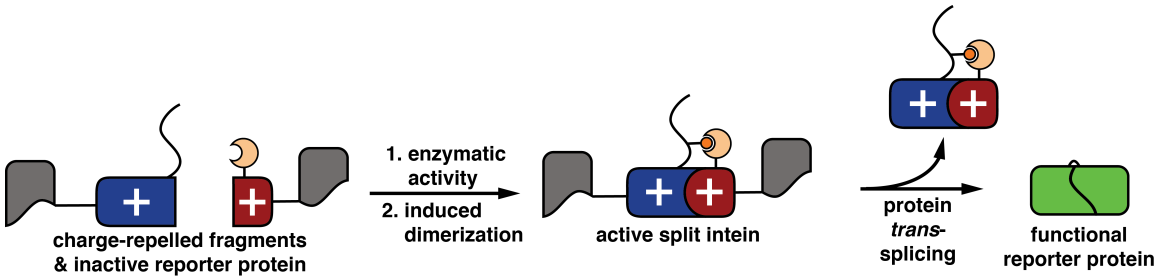


Figure 6.5. Conditional splicing with charge-repelled split intein fragments. In this design, an enzyme substrate sequence is fused to a mutant cationic NpuN, and a modular binding domain that recognizes the modified substrate is fused to wild-type cationic NpuC. In the presence of active enzyme, the N-intein is post-translationally modified and dimerization is induced. *Trans*-splicing ensues to create a functional reporter, such as a fluorescent protein.

6.3. Future directions for basic intein research

The work presented in this thesis addressed two goals: first, to understand the physiochemical basis for highly efficient protein *trans*-splicing by the split DnaE intein from *Nostoc puntiforme*, and second, to utilize this information to improve existing

intein-based protein engineering technologies and develop new technologies. *Trans*-splicing has two phases, fragment association and protein splicing. Based on the experiments outlined in Chapters 4 and 5, we now have a clearer picture of the binding process, and we know which molecular features of split inteins facilitate this process. From the experiments in Chapters 2 and 3, we now know that Npu is enhanced not only at the rate-limiting branch resolution step, but also in its initial N-extein activation step. Furthermore, comparison of related inteins demonstrated that Npu is not unique in its fast splicing kinetics and that there are conserved sequence motifs that confer this high efficiency. Ultimately, these findings have allowed us to address problems attendant to intein-based technologies, such as thioester isolation for Expressed Protein Ligation and C-extein dependence for protein *trans*-splicing applications. They should also pave the way for further protein engineering developments, as discussed earlier in this chapter. As it is clear that basic intein research can have downstream practical implications, future experiments should focus on furthering our basic understanding of intein chemistry and also intein biology.

6.3.1. Intein chemistry

In Chapter 2, we showed that specific mutations at non-catalytic residues in the Ssp intein improved that slow intein's activity (Figure 2.7).[123] While these effects were predicted by sequence-activity relationships between fast an slow-splicing inteins, the physical basis for these effects is not clear. Furthermore, none of these mutations restored full Npu-like activity in Ssp, and the additivity of their effects has not been tested. Recent experiments in our lab suggest that clusters of second- and third-shell residues

surrounding the active site motifs (including those identified by sequence homology in Chapter 2) can modulate the splicing step associated with those motifs. For example, mutations close to the Block B TXXH motif affect the initial N-to-S acyl shift reaction while mutations in the loop preceding the Block F HNF motif can affect branched intermediate resolution (as described above).* Future experiments should focus on determining the structural consequences of these mutations. Given that inteins do not appear to undergo large conformational changes throughout the splicing reaction, the observed functional effects may be due to changes in active site dynamics as a result of peripheral packing interactions. Thus, these structural studies will require a combination of techniques including X-ray crystallography, NMR spectroscopy, and maybe even molecular dynamics simulations. The ultimate goal of such studies will not only be to understand why certain inteins are fast but also to be able to predict the activity of new inteins based on their sequences.

6.3.2. Intein biology

Inteins typically assemble essential genes. For example, the cyanobacterial split inteins discussed throughout this thesis splice together fragments of the catalytic α-subunit of DNA polymerase III (DnaE).[19] Thus, it is puzzling why some inteins splice slowly while others are quick. There are two obvious possible explanations for this divergence in activity: (1) The conditions used for various analyses in the laboratory are not the optimal conditions for every intein. Indeed, cyanobacteria are an ecologically diverse class of microorganisms, and many species thrive in extreme conditions including

* Future experiments to identify the molecular determinants for fast splicing are being carried out by Adam Stevens.

high temperatures and salinity.[162] (2) An alternative explanation is that the evolutionary constraints on protein splicing kinetics differ from one organism to the next. Based on our sequence-activity analyses, it appears that the inefficient DnaE inteins have diverged from their more efficient relatives (Figure 2.5). It is possible that extremely rapid assembly of the DnaE gene product is necessary in some organisms such as *Nostoc punctiforme* (Npu) but this requirement is relaxed in others such as *Synechocystis* sp. PCC6803 (Ssp). Indeed, both of these organisms are mesophiles. The fact that these inteins assemble the only copy of the DNA polymerase required for genome replication raises the question of whether the evolution of fast or slow intein activity is dictated by the life cycle and replication rates of these organisms.

To address the questions described above, or the broader question "What is the biological role of inteins?", these auto-processing domains must be studied in their native contexts. Thus far, almost all basic intein research has been carried out *in vitro* or in a heterologous system, and the endogenous exteins have rarely been used as substrates. Future intein research should focus on trying to identify biological roles for intein beyond the biochemical reaction they carry out on model proteins. Given that many intein-containing organisms can be cultured and genetically manipulated (including budding yeast, many mycobacterial species, and even a number of cyanobacteria including Ssp and Npu[163]), these studies should be technically tractable. Some preliminary questions and experiments are listed below:

(1) Does an intein directly interact with biomolecules other than its covalently linked host protein? This may be addressed by a standard pull-down experiment coupled to mass spectrometry. As inteins have no known regulators in their natural environment,

any physical interactions they make in the cell would be biologically interesting and could even have technological implications.

(2) Does the protein splicing event have any physiological consequences? When the intein exists within an essential gene, it must excise itself for host organism survival. However, if the intein is genetically removed and the essential gene is appropriately fused, the organism should be viable. With this deletion, a synthetic lethality screen can be carried out to see if the effect of this deletion is buffered by any other biochemical pathways. An alternative approach to answer this question would be to profile mRNA levels in an intein-containing and intein-free strain of an organism to see if the presence of the protein splicing event and the resulting excised intein have any effect on the expression levels of particular genes.

(3) What are the splicing kinetics for different inteins *in vivo*? If epitope tags can be genetically introduced into the intein genes, a pulse-chase labeling experiment followed by immunoprecipitation of the inteins can be used to monitor auto-processing.

(4) If two orthologous inteins truly have different kinetics *in vivo*, are these kinetics important? A particularly intriguing experiment would be to generate a *Nostoc punctiforme* strain with the Ssp intein and a *Synechocystis* sp. PCC6803 strain with the Npu intein and look for altered phenotypes (either gross changes such as growth rate and viability or more subtle changes such as gene expression profiles).

(5) What are the structural consequences of intein insertion within its endogenous host protein? This question would primarily be addressed using structural and biophysical approaches *in vitro*. Currently, nothing is known about whether the host protein fragments actually interact with the intein nor if they can modulate its function

when compared to model exteins. Conversely, it is not clear what the structural effect of intein insertion is on the exteins. For example, can the intein affect and possibly aid in extein folding, thereby acting as an intramolecular chaperone? Prior to splicing, do the extein fragments have a different function? The latter question may also require *in vivo* experiments with a conditionally inactivatable intein mutant.

(6) Are split intein fragments differentially expressed? Given that many DnaE intein genes are on different chromosomal loci and thus do not share a promoter, it is possible that DnaE assembly is regulated by differential expression patterns of the fragments. Such as system could be biologically useful if one fragment has a biological function independent of the full-length DnaE protein. Alternatively, if the expression/maturation of one fragment is more challenging, the cells could slowly but constantly produce this piece and only expressed the simpler piece when needed to activate DnaE function. To test these hypotheses, cell growth would have to be synchronized to follow expression patterns by western blotting or mass spectrometry in a bulk culture. This may be feasible with cyanobacteria, which require cycles of light and dark to replicate.[164]

6.4. Outlook

The dearth of biologically relevant studies on inteins probably stems from the fact that inteins appear to be selfish genetic elements (especially when fused to homing endonuclease domains).[11] As such, biochemists and biologists have largely dismissed intein biology as impenetrable while focusing on the chemical properties of inteins and their immense utility for protein engineering.[8] Intein-based technologies have largely

been bolstered by mechanistic investigations into intein structure and activity, including the work described in this thesis. In the future, this basic research may include biological studies on intein function, and together with biochemical and biophysical approaches, will hopefully lead to the development of useful tools and the discovery of new biological processes.

Chapter 7: Methods*

7.1. Equipment

7.1.1. Size-exclusion chromatography

Size-exclusion chromatography (SEC) was carried out on a ÄKTA FPLC Purifier system from GE Healthcare. Analytical and preparative SEC was carried out on a Superdex 75 10/300 or Superdex 200 10/300 column and some preparative work was carried out on a Superdex 75 16/60 column. Multiple Angle Light Scattering (MALS) and Refractive Index (RI) measurements were performed using a Dawn Heleos II and Optilab T-rEX from Wyatt Technology Corporation (Santa Barbara, CA), respectively.

7.1.2. Reverse phase HPLC and mass spectrometry

Both analytical and semi-preparative RP-HPLC were performed on Hewlett-Packard 1100 and 1200 series instruments. Preparative HPLC was carried out on a Waters prep LC system comprised of a Waters 2545 Binary Gradient Module and a Waters 2489 UV detector. Analytical RP-HPLC was carried out on a C_{18} Vydac column (5 μm, 4.6 x 150 mm) at a flow rate of 1 mL/min. Semi-preparative RP-HPLC was carried out on a C_{18} Vydac 218TP152010 column (15-20 μM; 10 x 250 mm) at a flow rate of 4 mL/min. Preparative RP-HPLC was carried out on a C_{18} Vydac 218TP1022 column (10 μM; 22 x 250 mm) at a flow rate of 20 mL/min. All runs used 0.1 % TFA (trifluoroacetic acid) in water (solvent A) and 90 % acetonitrile in water with 0.1% TFA (solvent B). For analytical RP-HPLC runs, a two minute isocratic period in initial

* The experimental designs and details described in this chapter reflect my own work along with contributions from several collaborators. These co-contributors are acknowledged in footnotes on the first pages of research chapters 2-5.

conditions was followed by a 30 minutes linear gradient with increasing solvent B concentration. Electrospray ionization mass spectrometric analysis (ESI-MS) was performed on a Bruker Daltonics MicrOTOF-Q II mass spectrometer or a Sciex-API-100 single quadrupole spectrometer using positive ionization.

7.1.3. Spectroscopy

Steady-state fluorescence spectroscopy was carried out on a HORIBA Jobin Yvon FluroLog-3 spectrofluorometer. Stopped-flow fluorescence spectroscopy was performed on an Applied Photophysics SX20 stopped-flow spectrometer with excitation monochromaters and fluorescence detectors. Circular dichroism (CD) spectroscopy was carried out on an Applied Photophysics Chirascan CD spectrometer. NMR experiments were recorded on Varian Inova (600 MHz) and Bruker (600, 700, 800, 900 MHz) spectrometers equipped with cryogenic probes, and on Bruker (500, 800 MHz) spectrometers with room temperature probes.

7.1.4. Other equipment

In vivo intein activity assays were carried out on a VersaMax tunable microplate reader from Molecular Devices. Cells were lysed using an S-450D Branson Digital Sonifier or a French press. Western blots and coomassie-stained gels were imaged on a LI-COR Odyssey Infrared Imager. SYPRO ruby-stained gels were imaged using a Bio-Rad VersaDoc scanner. Fluorescent fluorescein-containing gels were imaged using the GE ImageQuant LAS 4000 imager. Densitometry analysis was done using the LI-COR Odyssey software or using ImageJ, developed at the National Institutes of Health.

Microwave-assisted solid-phase peptide synthesis (SPPS) was done on a CEM Liberty synthesizer.

7.2. Cloning

All plasmids for expression of recombinant proteins or for the bacterial Kan_R assay were prepared from existing plasmids using two procedures. To transfer genes between vectors, the restriction enzyme-free method, overlap-extension PCR with Phusion polymerase, was used as previously described.[165] To introduce point mutations, the Stratagene QuikChange site-directed mutagenesis kit was used with the manufacturer's instructions. For work with new DnaE intein sequences, codon-optimized synthetic genes were purchased for expression in *E. coli* and mammalian cells.

7.3. Preparation of peptides and proteins*

7.3.1. IntN fusions to large N-exteins

E. coli BL21(DE3) cells transformed with each N-intein plasmid were grown in 1 L of LB containing 100 μg/mL of ampicillin at 37°C until OD_{600} = 0.6. The cells were then cooled down to 18 °C, and expression was induced by addition of 0.5 mM IPTG for 16 hours at 18 °C. After harvesting the cells by centrifugation (10,500 rcf, 30 min), the cell pellets were transferred to 50 mL conical tubes with 5 mL of lysis buffer (50 mM phosphate, 300 mM NaCl, 5 mM imidazole, 2 mM BME, pH 8.0) and stored at -80°C. The cell pellets were thawed and resuspended by adding an additional 15 mL of lysis buffer supplemented with Complete protease inhibitor cocktail. Cells were lysed by

* RP-HPLC chromatographs showing purity of peptides/proteins and Tables of the expected and observed masses can be found at the end of the chapter.

sonication (35% amplitude, 8x 20 second pulses separated by 30 seconds on ice). The soluble fraction was recovered after centrifugation (35,000 rcf, 30 min) and mixed with 2 mL of Ni-NTA resin and incubated at 4°C for 30 minutes. After incubation, the slurry was loaded onto a fritted column. After discarding the flow-through, the column was washed with 5 column volumes (CV) of lysis buffer, 5 CV of wash buffer 1 (lysis buffer with 20 mM imidazole), and 3 CV of wash buffer 2 (lysis buffer with 50 mM imidazole). The protein was eluted with elution buffer (lysis buffer with 250 mM imidazole) in four 1.5 CV elution fractions. The wash and elution fractions were analyzed by SDS-PAGE.

After enrichment over the Ni-NTA column, the cleanest fractions were pooled and the proteins were purified by size exclusion chromatography. First, the protein was treated with 50 mM DTT or 10 mM TCEP for 30 minutes on ice and concentrated if necessary. Then, the protein was then injected on an S75 10/300 or S200 10./300 column and eluted over 1.35 CV in freshly prepared, degassed splicing buffer (100 mM phosphates, 150 mM NaCl, 1 mM DTT, 1 mM EDTA, pH 7.2). FPLC fractions were analyzed by SDS-PAGE, and the purest fractions were pooled and analyzed by analytical gel filtration, analytical RP-HPLC, and mass spectrometry. The concentration of pure proteins were determined by the Bradford assay and by the calculated extinction coefficient at 280 nm (ProtParam tool on the Expasy server).[166]

7.3.2. IntC fusions to large C-exteins

E. coli BL21(DE3) cells transformed with each C-intein plasmid were grown in 1 L of LB medium containing kanamycin (50 μg/mL) at 37 °C until OD_{600} = 0.6. Then, expression was induced by addition of 0.5 mM IPTG for 3 hours at 37 °C. The cells were

lysed, and the desired protein was enriched over Ni-NTA resin identically as for the ExtN-IntN proteins (above). The AvaC-SUMO, Csp(PCC8801)C-SUMO, and NpuC-eGFP proteins did not express well at 37 °C, so the proteins were expressed by induction at 18 °C for 16 hours. The cleanest fractions were pooled and dialyzed into TEV cleavage buffer (50 mM phosphate, 300 mM NaCl, 5 mM imidazole, 0.5 mM EDTA, 0.5 mM DTT, pH 8.0) then treated with His_6-tagged TEV protease overnight at room temperature. The cleavage was confirmed by RP-HPLC and ESI-MS, after which the reaction solution was incubated with Ni-NTA resin at room temperature for 30 min. The flow-through and two 1.5 CV washes with wash buffer 1 were collected and pooled. The protein was then concentrated if necessary, injected onto the S75 10/300 column, and eluted over 1.35 CV in freshly prepared, degassed splicing buffer (100 mM phosphates, 150 mM NaCl, 1 mM DTT, 1 mM EDTA, pH 7.2). FPLC fractions were analyzed by SDS-PAGE, and the purest fractions were pooled and analyzed by analytical gel filtration, analytical RP-HPLC, and mass spectrometry. The concentration of pure protein was determined by UV the Bradford assay and/or UV A_{280}.[166]

7.3.3. Contiguous DnaE intein fusions to various proteins

E. coli BL21(DE3) cells transformed with each protein-intein fusion plasmid were grown in 1 L of LB medium containing ampicillin (100 μg/mL) at 37 °C until OD_{600} = 0.6. Then, expression was induced by addition of 0.5 mM IPTG and incubation for 3 hours at 37 °C or incubation for 16 hours at 18 °C. All Ub fusions were expressed at 37 °C, the eGFP fusion was expressed at 18 °C, and the SH3, SH2, and $PARP_C$ fusions were expressed at both temperatures. After harvesting the cells by centrifugation (10,500 rcf,

30 min), the cell pellets were transferred to 50 mL conical tubes with 5 mL of lysis buffer (50 mM phosphate, 300 mM NaCl, 5 mM imidazole, No thiols, pH 8.0) and stored at -80°C. The cell pellets were thawed and resuspended by adding an additional 15 mL of lysis buffer supplemented with Complete protein inhibitor cocktail. Cells were lysed by sonication (35% amplitude, 8x 20 second pulses separated by 30 seconds on ice). The soluble fraction was recovered by centrifugation (35,000 rcf, 30 min). The soluble fraction was mixed with 2 mL of Ni-NTA resin and incubated at 4°C for 30 minutes. After incubation, the slurry was loaded onto a fritted column. After discarding the flow-through, the column was washed with 5 column volumes (CV) of lysis buffer without thiols, 5 CV of wash buffer 1 (lysis buffer with 20 mM imidazole, no thiols), and 3 CV of wash buffer 2 (lysis buffer with 50 mM imidazole, no thiols). The protein was eluted with elution buffer (lysis buffer with 250 mM imidazole, no thiols) in four 1.5 CV elution fractions. The wash and elution fractions were analyzed by SDS-PAGE with loading dye containing no thiols. The cleanest fractions were pooled and treated with 10 mM TCEP for 20 minutes on ice. Then, the solution was injected on an S75 or S200 10/300 column, and eluted over 1.35 CV in thiolysis buffer (100 mM phosphates, 150 mM NaCl, 1 mM EDTA, 1mM TCEP, pH 7.2). The FPLC fractions were analyzed by SDS-PAGE with loading dye containing no thiols, and the purest fractions were pooled and analyzed by analytical RP-HPLC and mass spectrometry. The concentration of pure protein was determined by UV A_{280nm}.[166]

7.3.4. CGK(Fl) peptide

Fmoc-based solid phase peptide synthesis (SPPS) was used to produce a peptide with the sequence H-Cys-Gly-Lys(Fluorescein)-NH_2 (CGK(Fl)). The peptide was

synthesized on Rink amide resin at a 0.2 mmol scale using Fmoc-Lys(Alloc)-OH, Fmoc-Gly-OH, and Boc-Cys(Trt)-OH as follows: 20% piperidine in DMF was used for Fmoc deprotection using a one minute equilibration of the resin followed by a 20 minute incubation. After Fmoc deprotection, amino acids were coupled using DIC/HOBt as activating agents. First, the amino acid (1.1 mmol) was dissolved in 50:50 DCM:DMF (2 mL) and was activated with DIC (1.0 mmol) and HOBt (1.2 mmol) at 0 °C for 15 minutes. The mixture was added to the N-terminally deprotected resin and coupled for 10 minutes at room temperature.

After the cysteine was coupled, the lysine side chain was deprotected by treatment with $Pd(Ph_3)_4$ (0.1 eq.) and phenylsilane (25 eq.) in dry DCM for 30 minutes. The peptidyl resin was washed with DCM (2 x 5 mL) and DMF (2 x 5 mL) followed by two washes with 0.5% DIPEA in DMF (v/v) and two washes with 0.5% sodium diethyldithiocarbamate trihydrate in DMF (w/v) to remove any remaining traces of the Pd catalyst. 5(6)-Carboxyfluorescein was then coupled to the lysine side chain using the DIC/HOBt activation method overnight at room temperature. Finally, the peptide was cleaved off the resin using 94% TFA, 1% TIS, 2.5% EDT, and 2.5% H_2O (6.5 mL) for one hour. After cleavage, roughly half of the TFA was evaporated under a stream of nitrogen. The crude peptide was precipitated with cold ether and washed with cold ether twice. Finally, the peptide was purified by RP-HPLC on C18 prep column over a 15-80% buffer B gradient in 40 minutes. The purified peptide was analyzed by analytical RP-HPLC and ESI-MS to confirm its identity. Note that no attempt was made to separately isolate the 5-carboxyfluorescein and 6-carboxyfluorescein conjugates, thus the peptide is a mixture of these two isomers.

7.3.5. hH2B(117-125)-K120Ac peptide

Fmoc-based solid phase peptide synthesis (SPPS) was used to produce a modified histone peptide with the sequence H-Cys-Val-Thr-Lys(Ac)-Tyr-Thr-Ser-Ala-Lys-OH. The peptide was synthesized on Wang resin preloaded with Fmoc-Lys(Boc)-OH at a 0.1 mmol scale on a CEM Liberty synthesizer. Fmoc deprotections were performed with 20% piperidine in DMF (5 mL) at 75 °C under 30 W microwave power for 30 s followed by an additional 3 min treatment. After Fmoc deprotection, amino acids were coupled using HBTU/HOBt/DIPEA as activating agents. First, the amino acid (0.5 mmol) dissolved in DMF (2.5 mL) was added to the reaction vessel followed by addition of HBTU (0.5 mmol) and HOBt (0.5 mmol) in DMF (1 mL) and DIPEA (1 mmol) in NMP (0.5 mL). Reaction mixture was heated to 75 °C under 20 W microwave power for 5 min. After peptide chain elongation was completed the peptide was cleaved off the resin using 94% TFA, 1% TIS, 2.5% EDT, and 2.5% H_2O for one hour at room temperature. After cleavage, half of the TFA was evaporated under a stream of nitrogen and the crude peptide precipitated with cold ethyl ether and washed with twice. Finally, the peptide was purified by RP-HPLC on C_{18} semi-preparative column over a 15-80% buffer B linear gradient over 30 minutes with an initial 2 min 15% B isocratic period. The purified peptide was analyzed by analytical RP-HPLC and ESI-MS to confirm its identity.

7.3.6. IntC column peptide for sEPL

The NpuC column peptide lacking the C-terminal cysteine was expressed with a C-terminal GyrA-H_6 tag and enriched over a Ni-NTA column. *In situ* intein cleavage and ligation to L-cysteine methyl ester (Cys(OMe)) was carried out by adding 100 mM

Cys(OMe), 5 mM TCEP, and adjusting the pH to 7.5. The reaction was allowed to proceed under room temperature for 24 hours. The reaction mixture was then injected onto a preparative C_{18} RP-HPLC column and purified over a linear gradient of 25-40% buffer B in 45 min. Pure fractions were combined and lyophilized to give Npu_C-AA-Cys(OMe) peptide. Peptide concentration was quantified using Ellman's reagent.

7.3.7. AEY-IntN constructs

E. coli BL21(DE3) cells transformed with each N-intein with a His_6-SUMO-AEY tag were grown in 2 L of LB containing 50 μg/mL of kanamycin at 37°C until OD_{600} = 0.6. The cells were then cooled down to 18 °C, and expression was induced by addition of 0.5 mM IPTG for 16 hours at 18 °C. After harvesting the cells by centrifugation (10,500 rcf, 30 min), the cell pellets were transferred to 50 mL conical tubes with 5 mL of lysis buffer (50 mM phosphate, 300 mM NaCl, 5 mM imidazole, 2 mM BME, pH 8.0) and stored at -80°C. The cell pellets were thawed and resuspended by adding an additional 35 mL of lysis buffer supplemented with Complete protease inhibitor cocktail. Cells were lysed by sonication (35% amplitude, 10x 20 second pulses separated by 30 seconds on ice).

While the N-intein could be isolated from the soluble lysate fraction, cleaner functional material could be extracted from the insoluble *E. coli* inclusion bodies in higher yields as follows. First, the lysate pellet was resuspended in 40 mL of Triton wash buffer (lysis buffer with 0.1% Triton X-100) and incubated at room temperature for 30 minutes. The Triton wash was centrifuged at 35,000 rcf for 30 minutes, and the supernatant was discarded. Next, the pellet was resuspended in 40 mL of lysis buffer

containing 6 M urea, and the mixture was incubated overnight at 4°C. The mixture was centrifuged at 35,000 rcf for 30 minutes, and then the supernatant was mixed with 4 mL of NiNTA resin. The Ni column was run identically as for the native purifications of other IntN fusions described above, except that every buffer had 6 M urea. Following enrichment over a Ni-NTA column, the protein was refolded into lysis buffer (without urea) by step-wise dialysis to remove the urea and excess imidazole at 4°C.

After enrichment over the Ni-NTA columns, the refolded proteins were treated with 10 mM TCEP and Ulp1 (SUMO-specific protease) overnight at room temperature. The cleavage was confirmed by RP-HPLC/MS, after which the reaction solution was incubated with Ni-NTA resin at room temperature for 30 min. The flow-through and two 1.5 CV washes with wash buffer 1 were collected and pooled. The protein was then concentrated to less than 10 mL, treated with 10 mM TCEP, injected onto the S75 16/60 gel filtration column, and eluted over 1.35 CV in freshly prepared, degassed splicing assay buffer (100 mM sodium phosphate, 150 mM NaCl, 1 mM EDTA, 1 mM DTT, pH 7.2) or NMR buffer (25 mM sodium phosphate, 100 mM NaCl, 1 mM DTT, pH 7.2). FPLC fractions were analyzed by SDS-PAGE, and the purest fractions were pooled and analyzed by analytical RP-HPLC and ESI-MS. The concentration of pure protein was determined by UV A_{280nm}.[166] The N-inteins could be stored at -80 °C after adding buffered glycerol to a final concentration of 20% and flash-freezing in liquid N_2.

7.3.8. Semi-synthetic IntC constructs with small C-exteins

E. coli BL21(DE3) cells transformed with an IntC-GyrAH_6 fusion plasmid were grown in 2 L of LB medium containing ampicillin (100 μg/mL) at 37 °C until OD_{600} =

0.6. Then, expression was induced by addition of 0.5 mM IPTG for 3 hours at 37 °C. After harvesting the cells by centrifugation (10,500 rcf, 30 min), the cell pellets were transferred to 50 mL conical tubes with 5 mL of lysis buffer (50 mM phosphate, 300 mM NaCl, 5 mM imidazole, no thiols, pH 8.0) and stored at -80°C. The cell pellets were resuspended by adding an additional 35 mL of lysis buffer supplemented with Complete protease inhibitor cocktail. Cells were lysed by sonication (35% amplitude, 10x 20 second pulses separated by 30 seconds on ice). The soluble fraction was recovered by centrifugation (35,000 rcf, 30 min). The soluble fraction was mixed with 4 mL of Ni-NTA resin and incubated at 4 °C for 30 minutes. After incubation, the slurry was loaded onto a fritted column. After discarding the flow-through, the column was washed with 5 column volumes (CV) of lysis buffer without thiols, 5 CV of wash buffer 1 (lysis buffer with 20 mM imidazole, no thiols), and 3 CV of wash buffer 2 (lysis buffer with 50 mM imidazole, no thiols). The protein was eluted with elution buffer (lysis buffer with 250 mM imidazole, no thiols) in four 1.5 CV elution fractions. The wash and elution fractions were analyzed by SDS-PAGE on TGX gels run in Tris/glycine/SDS running buffer.

The cleanest fractions containing the desired fusion protein were treated in one of two ways depending on the C-extein. For single amino acid C-exteins, the eluted protein from the Ni column (typically 20-30 mL at 50-100 μM) was treated with 10 mM TCEP and the Roche Complete protease inhibitor cocktail. Then the cysteine derivative (free carboxylate, methyl ester, carboxamide, or N-methyl amide) was added to the solution to a final concentration of 100 mM. The direct thiolysis/ligation reaction solution was kept at room temperature for 15-20 hours. For the di- and tri-peptide C-exteins (produced by solution or solid-phase synthesis, respectively, using standard protocols), the eluted

protein was first dialyzed at 4 °C against lysis buffer to remove excess imidazole. The dialyzed solution was then treated with 10 mM TCEP and the Roche Complete protease inhibitor cocktail. Next, 100 mM sodium 2-mercaptoethane sulfonate (MESNa) was added, followed immediately by 1-5 mM di- or tri-peptide. These one-pot thiolysis/ligation reaction solutions were kept at room temperature for 15-20 hours. For both reaction types, if the pH dropped after addition of all reagents, it was raised to pH 7.5 by addition of sodium hydroxide. Note that if the MESNa thiolysis and ligation reaction were not done in one pot, the C-terminal asparagine thioester would cyclize, then hydrolyze, resulting in a dead-end side product carboxylic acid.

After analysis by analytical RP-HPLC to confirm reaction completion, HPLC solvent B was added to achieve a final concentration of 10% B. Then, neat TFA was added to 0.5% to acidify the solution. Upon acidification, the cleaved GyrA-H_6 protein typically precipitated, but no C-intein adduct was lost. This solution was centrifuged then filtered for purification by preparative RP-HPLC. All C-intein constructs (typically 20-30 mL) were loaded on a C_{18} preparative column over a 10 minute isocratic phase in 10% B. Then the system was raised to the gradient starting conditions (20% or 25% B) over a 5 minute phase. Finally, the proteins were purified over a 20-35% B or 25-40% B gradient in 60 minutes. Fractions were analyzed by analytical RP-HPLC, and the purest fractions were lyophilized, then redissolved in water and pooled. Concentrations were determined by A_{280} using a calculated extinction coefficient.[166] Purity of the constructs was assessed by analytical RP-HPLC, and their identities were confirmed by ESI-MS. Isolated C-inteins could be stored at -80 °C in water and freeze-thawed multiple times.

7.3.9. Three-piece ligation constructs with orthogonal inteins

These proteins were expressed and purified similarly to the IntN constructs with large exteins except for the last SEC step. Briefly, proteins were expressed in *E. coli* at 18 °C, cells were lysed by sonication, and the proteins were extracted from the soluble fraction of the cell lysate. After enrichment over Ni-NTA resin, the Ni column fractions were analyzed by SDS-PAGE, and the cleanest fractions were kept for three-piece ligations with no further purification.

7.3.10. NpuN lobe constructs

These proteins were expressed and purified similarly to the AEY-IntN constructs, except that the proteins were enriched directly from the soluble fraction of the cell lysate as H_6-SUMO fusions. The H_6-SUMO tags were cleaved off with Ulp1, and the reactions was clarified by reverse Ni affinity chromatography. The proteins were further purified by size exclusion chromatography on a S75 column in NMR buffer (25 mM sodium phosphates, 100 mM NaCl, 1 mM DTT, pH 6.5). The identity of the products were confirmed by ESI-MS, and the purities were assessed by analytical RP-HPLC.

7.3.11. Segmentally labeled NpuN

For segmental labeling of NpuN, the N-intein was split between Y58 and C59. NpuN residues 1-58 with the C1A mutation and the AEY N-extein were expressed in BL21(DE3) cells with a C-terminal GyrA-H_6 tag. The intein fusion was enriched over Ni-NTA resin and the tag was cleaved off by overnight thiolysis with 100 mM MESNa at room temperature in a phosphate buffered saline at pH 8.0. The resulting reaction

mixture, containing the AEY-NpuN(C1A-58)-MES α-thioester and cleaved GyrA, was treated with 1% trifluoroacetic acid (TFA) to precipitate the proteins. The proteins were redissolved in ligation buffer (6 M guanidine hydrochloride, 200 mM sodium phosphates, 100 mM MESNa, pH 7.5) and purified by SEC on an S75 column in ligation buffer (Figure 7.1a-c). The second fragment, NpuN(59-102), was expressed with an N-terminal H_6-SUMO tag and enriched identically as isolated $NpuN_2$ described above. The SUMO tag was cleaved off to yield a free N-terminal C59, and the cleavage buffer was supplemented with 5 mM L-cysteine to scavenge anything that might react with the liberated C59. The H_6-SUMO tag, Ulp1, and un-reacted material were removed by reverse Ni affinity chromatography. The protein was denatured by dialysis into ligation buffer then purified by SEC on an S75 column in ligation buffer (Figure 7.1d-f).

For EPL, the purified thioester and N-terminal cysteine-containing fragments were mixed at a ratio of 1:1.1 molar equivalents and treated with 10 mM TCEP. The mixture was concentrated to approximately 1 mM fragment concentrations and the reaction was allowed to proceed overnight at room temperature (Figure 7.1g-h). The product was purified by semi-preparative RP-HPLC, then refolded by stepwise dialysis into NMR buffer (25 mM sodium phosphates, 100 mM NaCl, 1 mM DTT, pH 6.5). The refolded N-intein was further purified by SEC on an S75 column in NMR buffer. The purity of the refolded proteins was analyzed by RP-HPLC and their identities were confirmed by ESI-MS (Figure 7.1i-j).

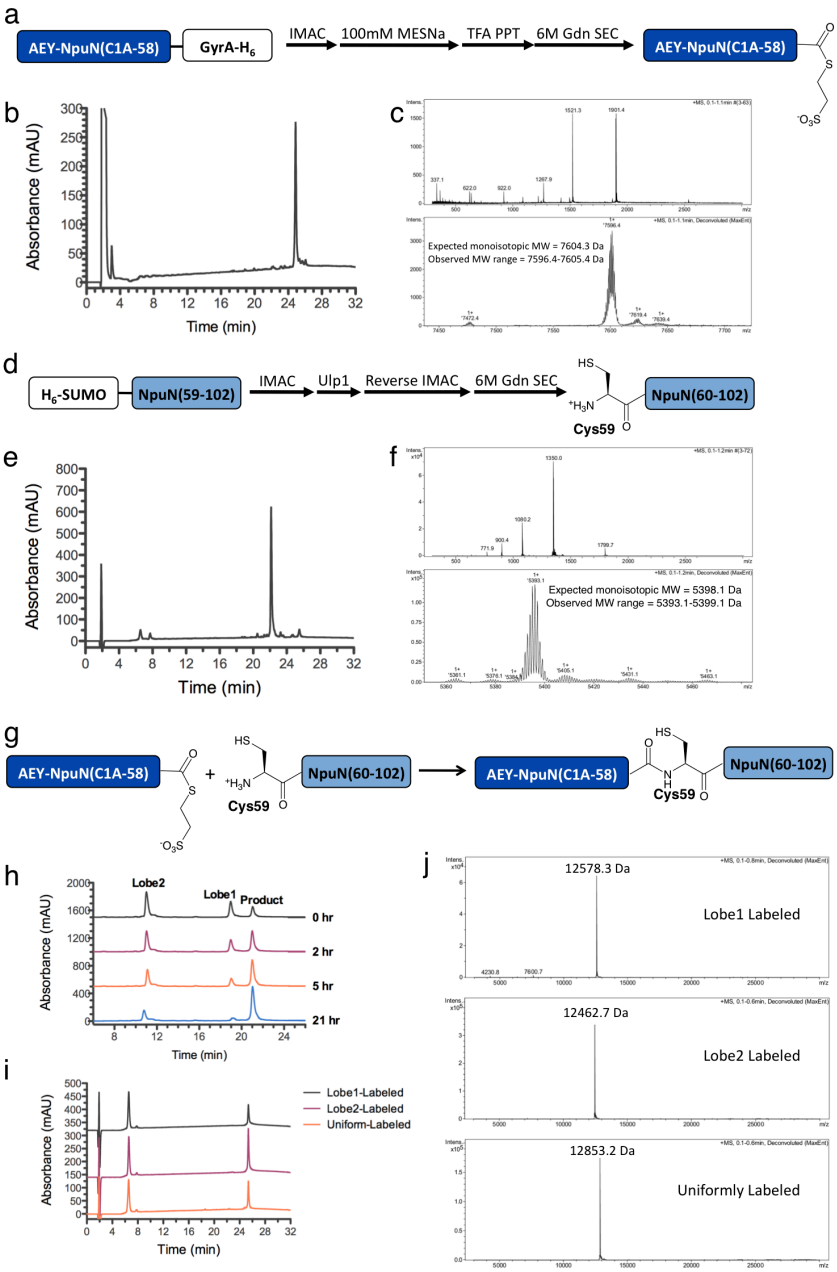


Figure 7.1. Segmental labeling of NpuN. a. Scheme for $NpuN_1$ thioester purification. **b.** RP-HPLC chromatogram and **c.** mass spectrum of ^{13}C- and ^{15}N-labeled $NpuN_1$ thioester. **d.** Scheme for purification of $NpuN_2$ with a native N-terminal cysteine. **e.** RP-HPLC chromatogram and **f.** mass spectrum of ^{13}C- and ^{15}N-labeled $NpuN_2$ with a native N-terminal cysteine. **g.** Ligation reaction scheme. **h.** RP-HPLC analysis of ligation reaction progress. **i.** RP-HPLC chromatograms of pure ligated products with either labeling scheme compared with a chromatogram of uniformly labeled sample. Peaks betwee 6 and 8 minutes are from DTT in the sample buffer. **j.** Deconvoluted mass spectra of purified proteins. Proteins were run on an analytical C18 column over a 0-73% B gradient in 30 minutes, except panel h which was over a 35-70% B gradient.

7.3.12. Isotopically enriched samples expressed in M9 medium

For all proteins studied by NMR, isotopically ^{15}N- and ^{13}C-enriched versions were prepared identically as the unlabeled versions with the exception that they were expressed in *E. coli* cultured in minimal M9 medium. 1 L of M9 medium was prepared as follows. First, 6.7 g of Na_2HPO_4 (anhydrous), 3.0 g of KH_2PO_4 (anhydrous), and 0.5 g NaCl were dissolved in 1 L of Milli-Q water. The buffer pH was adjusted to 7.2. 100 μL of 1 M $CaCl_2$ were added to this solution (some precipitate forms) and the buffer was autoclaved. Before inoculation, the following autoclaved components were added: (1) 1 g of $^{15}NH_4Cl$, (2) 5 g of D-glucose or 2 g of uniformly ^{13}C-labeled D-glucose (3) 2 mL of 1 M $MgSO_4$, (4) 10 mL of 100x Kao and Michayluk vitamin solution (Sigma cat. #K3129), (5) 1 mL of metal solution 1 (50mM H_3BO_3, 0.2mM $CoCl_2$, 1mM $CuSO_4$, 1mM $MnCl_2$, 0.001mM Na_2MoO_4, 2mM $ZnCl_2$), and (6) 1 mL of metal solution 2 (1 mM $FeSO_4$). Overnight cultures (10 mL for every liter of expression culture) were grown as usual in LB medium, then the dense starter culture was centrifuged to pellet the bacteria. The LB supernatant was decanted, the pellet was resuspended in 10 mL of M9, and the resuspension was added to the 1 L of M9 for expression. The expressions and purifications were carried out normally from this point onward.

7.4. Intein activity assays

7.4.1. In vivo kanamycin resisitance assays

96-well plate assay. Intein activity-coupled kanamycin resistance (Kan_R) assays were conducted in 96-well plate format as previously described.[92] Typically, plasmids were transformed into 15 μL of sub-cloning efficiency DH5α cells by heat shock, and the

transformed cells were grown for 18 hours at 37°C in 3 mL of LB medium with 100 μg/mL of ampicillin (LB/amp). The over-night cultures were diluted 250-fold into LB/amp solutions containing 8 different kanamycin concentrations (150 μL per culture). The cells were grown at 30°C on a 96-well plate, monitoring optical density (OD) at 650 nm every 5 minutes for 24 hours while shaking for one minute preceding each measurement. The endpoint of this growth curve (typically in the stationary phase) was plotted as a function of kanamycin concentration to visualize the dose-response relationship and these curves were fit to a variable-slope dose-response equation to determine IC_{50} values.

$$OD_{Obs} = OD_{Min} + \frac{\left(OD_{Max} - OD_{Min}\right)}{1 + 10^{\left[\left(\log IC_{50} - \log[Kan]\right)\cdot HillSlope\right]}}$$

In each regression analysis, typically three or four independent dose response curves were collectively fit to the equation above using the GraphPad Prism software. In each fit, OD_{Min} was fixed to the background absorbance at 650 nm, and all other parameters were allowed to vary. The reported error bars for the IC_{50} bar graphs represent the standard error in the best-fit IC_{50} value from three or four collectively fit dose-response curves.

Western blotting of cell lysates. For the western blot analyses, DH5α cells were transformed with the assay plasmids identically as for the 96-well plate setup and grown for 18 hours at 37°C while shaking. The overnight cultures were used to inoculate 3 mL of fresh LB/amp at a 1:300 dilution, and the cells were incubated at 30°C for 24 hours. The ODs of the 30°C cultures were measured at 650nm to assess relative bacterial levels, then 150 μL of each culture was transferred to an Eppendorf tube and centrifuged at 17,000 rcf for 2 minutes. The supernatant was aspirated off, and the cell pellets were

resuspended/lysed in ~200 μL of 2x SDS gel loading dye containing 4% BME (the resuspension volumes were varied slightly to normalize for differences in OD). The samples were boiled for 10 minutes, then centrifuged at 17,000 rcf for 1 minute. Each sample (5 μL) was loaded onto a 12% Bis-Tris gel and run in MES-SDS running buffer. The proteins were transferred to PVDF membrane in Towbin transfer buffer (25 mM Tris, 192 mM glycine, 15% methanol) at 100 V for 90 minutes. Membranes were blocked with 4% milk in TBST, then the primary antibody (α-myc, 1:5000) and secondary antibody (LI-COR mouse 800, 1:15,000) were sequentially applied in 4% milk in TBST. The blots were imaged using the LI-COR Odyssey scanner.

7.4.2. SDS-PAGE protein splicing assays

Assay procedure. For a typical assay, individual protein stock solutions of each relevant construct were prepared in filtered splicing buffer (100 mM phosphate, 150 mM NaCl, 1 mM DTT, 1 mM EDTA, pH 7.2) at 2 times the final concentration for two component reactions and three or four times the final concentration for three or four component reactions (e.g. 2.0 μM stock solution for a 1.0 μM standard *trans*-splicing reaction). 1 mM TCEP was added (from a pH-neutralized 100 mM stock solution) to each protein solution, and the proteins were incubated for 5 min at the appropriate reaction temperature. To initiate a reaction, the components were mixed at equal volumes. A typical reaction volume was less than 500 μL and was carried out in an Eppendorf tube on a heat block. During the reaction, 10-30 μL aliquots of the reaction solution were removed at the desired time points and quenched in 2x, 4x, or 6x concentrated SDS gel loading dye on ice to afford a final quenched solution with 40 mM

Tris (~pH 7.0), 10% (v/v) glycerol, 1% (w/v) SDS, 0.02% (w/v) bromophenol blue, and 2% (v/v) BME. For each reaction, an artificial zero time point was taken by mixing equivalent amounts of starting materials directly into the quencher solution. Samples were boiled for 10 minutes then centrifuged at 17,000 rcf for 1 minute. Aliquots of starting materials and time points were loaded onto Bis-Tris gels and run in MES-SDS running buffer. The gels were Coomassie-stained then imaged using the LI-COR Odyssey scanner or stained with SYPRO Ruby (for 0.05 μM and 0.10 μM reactions) and imaged using a VersaDoc scanner.

Note that for the reactions with a CGN C-extein sequence described in Chapter 3, no BME was used in the quencher solution. Furthermore, before boiling the samples, each sample was treated with 1 μL of 2 N HCl. After boiling and cooling the samples, they were treated with 1 μL of 2 N NaOH. This procedure reduced undesired hydrolysis or thiolysis of the branched intermediate during sample preparation.

Determination of kinetic parameters. To determine reaction rates, each lane of a gel was analyzed using the LI-COR Odyssey quantification function or ImageJ. Given the close proximity of the starting material bands in the Ub-SUMO *trans*-splicing reactions, these bands were typically integrated together. To normalize for loading error, the integrated intensity of each band in a lane was expressed as a fraction intensity of the total band intensity in that lane (which remained relatively constant between lanes). These normalized intensities were plotted as a function of time, and data from three independent reactions were collectively fit to first-order rate equations using the GraphPad Prism software:

For reactant depletion: $$Y = S \cdot \left(e^{-k_{obs} \cdot t}\right) + Z$$

For product formation: $$Y = Y_{\max} \cdot \left(1 - e^{-k_{obs} \cdot t}\right)$$

Y is the fractional intensity of a species, t is time in minutes, S is a scaling factor for reactant depletion (allowed to vary), Z indicates the fraction of reactant remaining at the reaction endpoint (allowed to vary), Y_{max} is a scaling factor for product formation, and k_{obs} is the observed first-order rate constant for the splicing reaction (allowed to vary). Half-lives were calculated from the best-fit value for the first-order rate constant:

$$t_{1/2} = \frac{\ln 2}{k_{obs}}$$

For reactions with no detectable side product formation, the rate of product and IntN formation were consistent with the rate of starting material depletion..

Alternatively, to extract second-order rate constants when intein fragment binding was significantly perturbed, as described in Chapter 4, the following equations were used:

For reactant depletion: $$Y = S \cdot \left(\frac{1}{1 + \left(\left[Intein\right]_0 \cdot k_{rxn} \cdot t\right)}\right) + Z$$

For product formation: $$Y = Y_{\max} \cdot \left(1 - \frac{1}{1 + \left(\left[Intein\right]_0 \cdot k_{rxn} \cdot t\right)}\right)$$

Y is the fractional intensity of a species, t is time in minutes, $[Intein]_0$ is the initial intein concentration (fixed based on reaction conditions), S is a scaling factor for reactant depletion (allowed to vary), Z indicates the fraction of product remaining at the reaction endpoint (allowed to vary), Y_{max} is a scaling factor for product formation, and k_{rxn} is the second-order rate constant for the splicing reaction (allowed to vary). Initial half-lives were calculated from the best-fit value for the second-order rate constant:

$$t_{1/2} = \frac{1}{\left[Intein\right]_0 \cdot k_{rxn}}$$

For all reactions, individual product formation data sets were normalized based on

scaling factors, and 3-4 normalized data sets from different reactions were globally fit to the same equation given above. It is noteworthy that for all reactions in Chapter 4, a first-order and second-order fit gave similar half-lives (initial half-lives for second-order fits), however regression analysis with a second-order rate equation gave better fits.

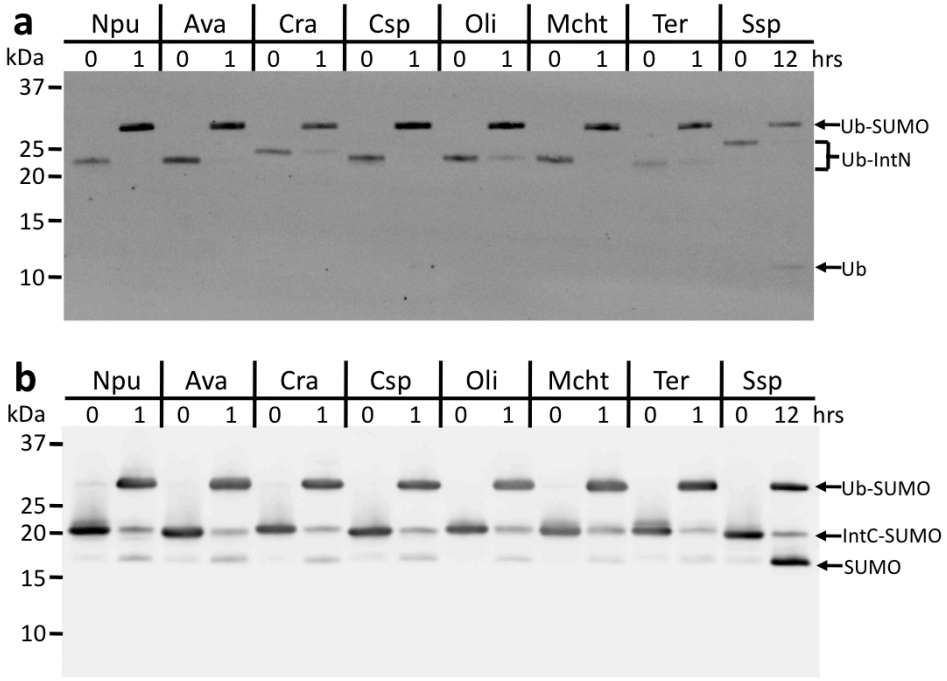


Figure 7.2. Western blots to identify bands in SDS-PAGE splicing assays. a. α-His_6 blot to detect the His_6-Ub N-extein. **b.** α-HA blot to detect the SUMO-HA C-extein.

Western blotting of reaction time points. For the two-component splicing assays described in Chapter 2 and the four-component competition assays described in Chapter 4, western blotting was used to confirm band identities and quantitatively follow the reaction, respectively. For the reactions in Chapter 2, western blots of the zero time point and reaction endpoint were carried out to confirm the identities of the observed bands. The quenched time points from the reactions described above were loaded onto 12% Bis-Tris gels (5 μL per sample, two identical gels) and run in MES-SDS running buffer. The resolved proteins were transferred from the gel onto PVDF membrane in CAPS transfer

buffer (10 mM *N*-cyclohexyl-3-aminopropanesulfonic acid, 10% (v/v) methanol, pH 10.5) at 100 V for 60 minutes. Membranes were blocked with LI-COR Blocking Buffer, then the primary antibody (Millipore α-His_6, 1:3000, or Covance α-HA, 1:25,000) was applied in LI-COR Blocking Buffer. The secondary antibody (LI-COR mouse 800, 1:15,000) was applied in 4% milk in TBST. The blots were imaged using the LI-COR Odyssey scanner. For the competition assays, the same procedure was used to analyze all time points in the reaction, blotting only for the HA epitope on the C-exteins.

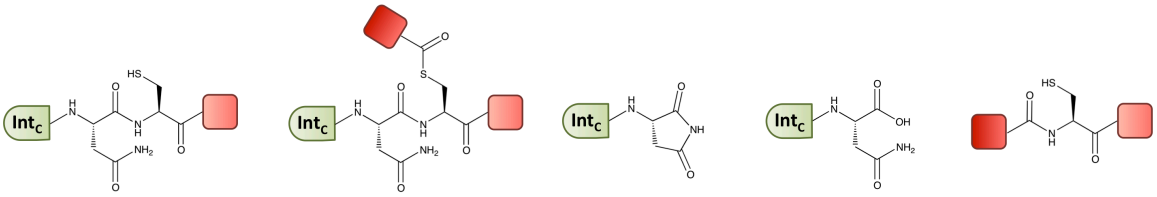


Figure 7.3. Scheme of observable species (1-5) in kinetic assays in Chapter 3. Molecule **1** is found before association of the split intein fragments, in the precursor complex, and in the linear thioester intermediate. Molecule **2** is unique to the branched intermediate state of the *trans*-splicing reaction. Molecules **3-5** are products of *trans*-splicing. Specifically, molecule **3** comes from the excised intein complex, however it slowly hydrolyzes to **4**. Spliced product **5** initially exists in a thioester form immediately upon intein excision, however this is not detectable, as it rapidly isomerizes to the amide form sown in the scheme above.

7.4.3. RP-HPLC and ESI-MS protein splicing assays

Assay procedure. Prior to any splicing assay, the N-intein solutions were dialyzed against splicing assay buffer (100 mM sodium phosphates, 150 mM NaCl, 1 mM EDTA, pH 7.2) containing no thiols overnight at 4 °C. If thiols were present, substantial N-extein cleavage could be observed for reactions with a slow k_3, however 1 mM DTT had no impact on reactions with $k_3 > 1.5 \times 10^{-3}$ s^{-1}. Note that for reaction 15, even in the absence of thiols, some N-extein cleavage was observed due to the extremely slow k_3. N-extein cleavage is characterized by a re-emergence of the starting material, species **1** in Figure

7.3). After removing thiols, N-inteins were diluted to 15 μM in assay buffer, C-inteins were diluted to 10 μM, and both solutions were treated with 2 mM TCEP. The solutions were incubated at 30 °C for 10 minutes, then reactions were initiated by mixing equal volumes of N- and C-inteins and continuing to incubate at 30 °C. During the reaction, aliquots of the solution were removed and mixed 3:1 (v/v) with quenching solution (8 M guanidine hydrochloride and 4% trifluoroacetic acid).

For RP-HPLC analysis, reactions were typically carried out on a 1.3-1.4 mL scale, where 100 μL were removed at each time point and mixed with 33 μL of quenching solution. 100 μL of the quenched solutions were separated 0-73% B gradient in 30 minutes on a C_{18} analytical column, recording absorbance at 214 nm. The major peaks were collected and identified by ESI-MS. For direct ESI-MS analyses, reactions were typically carried out on a 600 μL scale, where 30 μL were removed at each time point and mixed with 10 μL of quenching solution. 20 μL of the quenched solutions were desalted using Millipore C_{18} Zip-Tips. The eluates from the Zip-Tips were diluted 20-fold in 50% acetonitrile in water with 0.1% formic acid and loaded on the mass spectrometer by direct infusion. The complex mixture of multiply-charged states of each species were deconvoluted using the Maximum Entropy algorithm (Spectrum Square Associates, Ithaca, NY) into spectra depicting a well-defined mixture of singly-charged species.

Determination of kinetic parameters. The RP-HPLC and ESI-MS data were quantified to yield reaction progress curves. First, using the manufacturer's analytical software, the peak areas for all relevant chromatographic or mass spectrometric peaks were calculated. For RP-HPLC, the relevant peaks came from species **1**-**5**, and for ESI-MS, the relevant peaks came from species **1**-**4** (Figure 7.3). Then, for each time point in a

reaction, the area of an individual peak was expressed as a fraction of the total peak area. Reaction progress curves for each species were generated by expressing this normalized intensity as a function of time. Note that species **3**-**5** are all products that irreversibly form after branched intermediate resolution (Figure 3.1). Since **3** and **5** are direct products of *trans*-splicing, and **3** converts into **4** over time:

$$\frac{d[3+4+5]}{dt} = \frac{d[3+4]}{dt} = \frac{d[5]}{dt}$$

To account for changes in absorbance, in the RP-HPLC reactions, the peak area of the product is considered to be the total peak areas **3** + **4** + **5**. On the other hand for ESI-MS, based on size and composition, we assume that species **1**-**4** all have similar ionization efficiencies by ESI-MS, whereas **5** is a small molecule with dramatically different ionization properties. Thus the product from ESI-MS reactions is simply **3** + **4**. Note that N-intein species (AEY-IntN and IntN) do not separate under any RP-HPLC conditions tested (not shown), thus they were not used for any quantitative analysis. It is noteworthy that the two assay formats gave extremely similar quantitative results (Figure 7.4).

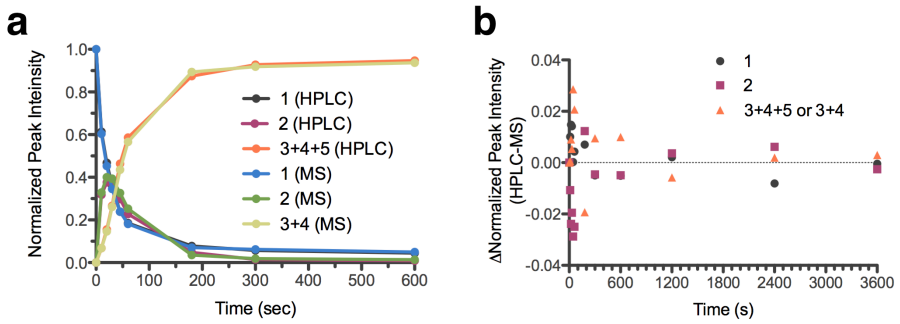


Figure 7.4 Comparison of RP-HPLC and ESI-MS splicing assay in Chapter 3. a. Analysis and quantification of the same samples from one iteration of reaction 1 in Chapter 3 by RP-HPLC (first three reaction curves) and ESI-MS (second three reaction curves). **b.** Difference between the normalized peak intensities for the RP-HPLC and ESI-MS reactions for each species in the kinetic model. Note that the maximal error is roughly 3% and is only that large for the first few time points.

Our data were fit to the simplified three-state kinetic model shown and below:

$$\underset{\mathbf{1}}{\text{Starting Mat. / Linear Int.}} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \underset{\mathbf{2}}{\text{Branched Int.}} \xrightarrow{k_3} \underset{\mathbf{3} + \mathbf{4}\ (+\ \mathbf{5})}{\text{Product}}$$

For each reaction, the normalized reaction progress curves for **1**, **2**, and **3** + **4** (+ **5**) were collectively fit to a system of equations that are the analytical solution to the coupled differential rate equations for those species:

$$p = k_1 + k_2 + k_3$$

$$q = \sqrt{p^2 - 4(k_1 k_3)}$$

$$a = \tfrac{1}{2}(p + q)$$

$$b = \tfrac{1}{2}(p - q)$$

$$[S/L](t) = [S/L]_0 \left[\left(\frac{k_1(a - k_3)}{a(a - b)} \right) e^{-at} + \left(\frac{k_1(k_3 - b)}{b(a - b)} \right) e^{-bt} \right]$$

$$[B](t) = [S/L]_0 \left[\left(\frac{-k_1 a}{a(a - b)} \right) e^{-at} + \left(\frac{k_1 b}{b(a - b)} \right) e^{-bt} \right]$$

$$[P](t) = [S/L]_0 \left[\left(\frac{k_1 k_3}{ab} \right) + \left(\frac{k_1 k_3}{a(a - b)} \right) e^{-at} - \left(\frac{k_1 k_3}{b(a - b)} \right) e^{-bt} \right]$$

In these equations, p, q, a, and b are algebraic combinations of rate constants k_1, k_2, and k_3 that simplify expression of the rate equations. [*S*/*L*] is the normalized intensity of the **S**tarting Material / **L**inear intermediate (**1**), [*B*] is the normalized intensity of the **B**ranched Intermediate (**2**), and [*P*] is the normalized intensity of the **P**roducts (**3** + **4** + **5** for RP-HPLC analysis and **3** + **4** for ESI-MS analysis). The variable *t* is the reaction time in seconds, and $[S/L]_0$ can be considered a normalization factor that represents the extent of the reaction (the fraction of starting material that gets converted to product).

The data were fit to the equations above using the multiple fit function in the Pro Fit data analysis software (QuantumSoft, Switzerland). In these fits, k_1, k_2, k_3, and $[S/L]_0$

were allowed to vary, and best-fit values were identified using the Levenberg-Marquardt algorithm. For all reactions except reaction 15 (which had significant N-extein cleavage), $[S/L]_0$ was between 0.81 and 0.95. This fitting process was independently carried out for individual replicates, and the best-fit values from three or more replicates were averaged. These averages and the standard deviation between these replicates are given in Chapter 3. Finally, the overall rate of splicing (k_{splice}) was determined by treating product formation, **3** + **4** (+ **5**), as a single-step first order reaction and fitting those reaction progress curves to the following equation using GraphPad Prism:

$$[P](t) = [P]_{\max} \cdot \left(1 - e^{-k_{splice} \cdot t}\right)$$

where $[P]_{max}$ is analogous to $[S/L]_0$ from the previous equations. The average and standard deviation for k_{splice} from three independently fit reactions can be found in Table 1 of the main text.

There were five exceptions to our general curve-fitting protocol. For reactions 4 and 5, branched intermediate resolution was not observed. Thus, the reactions were treated as an equilibrium between two states and the curves for **1** and **2** were simultaneously fit to these equations:

$$[S/L](t) = \frac{1}{\left(k_1 + k_2\right)} \cdot \left(k_2 + k_1 e^{-(k_1 + k_2)t}\right)$$

$$[B](t) = \frac{k_1}{\left(k_1 + k_2\right)} \cdot \left(1 - e^{-(k_1 + k_2)t}\right)$$

For reaction 15, N-extein cleavage caused a re-emergence of starting material **1**, thus the equations for the three-state model could not be readily applied for global fitting. Since BI resolution was extraordinarily slow for this reaction, the reaction was comprised of an initial pre-equilibrium step and a product formation step that were de-coupled. Thus, the

data for **1** and **2** from the first 10 minutes of the reaction curve were fit to the two-state model described above to extract k_1 and k_2. The data for product accumulation (**3** + **4**) over the course of the entire reaction, were fit to the single-step first order rate equation to extract k_{splice}, which we assumed to be equal to k_3. The last two exceptions were reactions 3 and 13, reactions with the C1A mutation. For these reaction, only C-extein cleavage was observed, thus they were treated as single-step first-order reactions. For reaction 3 the curve for **3** + **4** + $CFN(NH_2)$ was fit to the single-step first-order rate equation. For reaction 13, the curve for **3** + **4** was fit to the same equation. Since reaction 13 was extremely slow and did not plateau after a few days, $[P]_{max}$ was constrained to 0.95, which was the plateau value for the analogous reaction 3.

7.5. Additional intein activity and technology experiments

7.5.1. Observation of the linear thioester intermediate by RP-HPLC

For Npu, Ava, and Mcht fusions to ubiquitin, three peaks were visible for the purified protein when directly injected onto a C18 RP-HPLC column from a neutral buffer. These peaks all had the same mass of the desired protein. When diluted 20-fold in H_2O containing 0.1% TFA (pH 2) and incubated for at least two hours at room temperature, the first two peaks merged into the third peak. The same observation could not be made for MxeGyrA under identical conditions. To further confirm that we were observing an equilibrium between the precursor amide and linear thioester, the Ub-NpuDnaE protein was diluted 20-fold in thiolysis buffer containing 1% SDS. Before boiling, the two major peaks were visible. After boiling for 10 minutes, when the protein was unfolded, the first major peak partially converged into the second major peak,

suggesting that the latter was the amide, which should be more stable in the unfolded intein. Additional evidence that the three peaks were in equilibrium came from pH titrations. The protein was diluted 20-fold into citric acid/phosphate buffers ranging from pH 2 to pH 8, incubated at room temperature for 3-4 hours, then analyzed by HPLC over a 30-73% B gradient in 30 minutes. The relative abundance of the three species was modulated and showed a bell-shaped pH dependence, similar to the activities of enzymes containing multiple ionizable functional groups in their active sites.

In addition to observing the desired protein mass from all three observed HPLC peaks, we also observed the presence of a -18 Da species in the first two peaks. This mass change is characteristic of a dehydration reaction, and such a reaction has been previously reported by Mootz *et. al.* for a mutant form of the SspDnaB intein that cannot efficiently catalyze the initial N-to-S acyl shift.[167] Specifically, the tetrahedral intermediate of the forward and reverse acylation reactions can undergo acid-catalyzed dehydration to yield a thiazoline side product. As our DnaE intein constructs are fully active, this dehydrated species is most likely due to acidification of samples for mass spectrometric analysis.

7.5.2. Thiolysis/ligation with contiguous DnaE inteins

For each Ub-Intein fusion protein, four reactions were carried on a 100 μL scale at 30 °C. In the first reaction to monitor background hydrolysis, the fusion protein (50 μM) was incubated in thiolysis buffer (100 mM phosphate, 150 mM NaCl, 1 mM EDTA, 1mM TCEP, pH 7.2) supplemented with freshly added TCEP (an additional 5 mM). In the second and third reactions, the protein was incubated identically as for the first reaction, except that each reaction had either 100 mM MESNa or 1 mM CGK-

Fluorescein. In the fourth reaction, both MESNa and the peptide were added. At various time points, 5 μL of reaction solution were removed and quenched in 30 μL 2x SDS loading dye containing no thiols. As time points were collected, they were stored at -20 °C until the end of the reaction. After the reaction, the 35 μL quenched time points were thawed, treated with 1 μL of a 1 M TCEP stock solution, boiled for 10 minutes, and centrifuged at 17,000 rcf for 1 minute. Time points (5 μL) were loaded onto 12% Bis-Tris gels and run in MES-SDS running buffer. The gels were first imaged on a fluorescence imager to visualize the Ub-CGK(Fl) ligation product. Then the gels were coomassie-stained and imaged using the LI-COR Odyssey scanner. In addition, the reaction endpoints were quenched by 20-fold dilution in H_2O with 0.1% TFA and injected on an analytical C_{18} RP-HPLC column. The mixture was separated over a 2 minute isocratic phase in 0% B followed by a 0-73% B linear gradient in 30 minutes. The major peaks were collected and analyzed by MS. Thiolysis of other contiguous DnaE intein fusions was carried out identically.

7.5.3. Preparation of the IntC column for sEPL

The Npu_C-AA-Cys(OMe) peptide was dissolved in 2 CV coupling buffer (50 mM Tris, 5 mM EDTA, pH 8.5) and treated with 25 mM TCEP from a 1 M stock for 15 min. The peptide solution was then added to 1 CV of Agarose SulfoLink resin (loading: 18.4 μmol iodoacetyl groups/mL of resin) in a small fritted column, and incubated for 15 min on a nutator, followed by 30 min standing at room temperature. The column flow-through was collected and the column washed twice with 2 CV of coupling buffer. Flow-through and wash fractions were combined and analyzed by RP-HPLC. Comparison with

a sample of the input solution was used to determine NpuC-AA loading. Un-reacted iodoacetyl groups on the resin were blocked by treatment with 50 mM Cys(OMe) in coupling buffer for 15 min on a nutator, followed by 30 min standing at room temperature. The column was washed twice with 1 CV of coupling buffer, 2 CV of 1 M NaCl and finally 2 CV of water. Int_C-columns were stored in 100 mM phosphate, 150 mM NaCl, 1 mM EDTA, 0.05% NaN_3, pH 7.2 at 4 °C for up to 2 weeks.

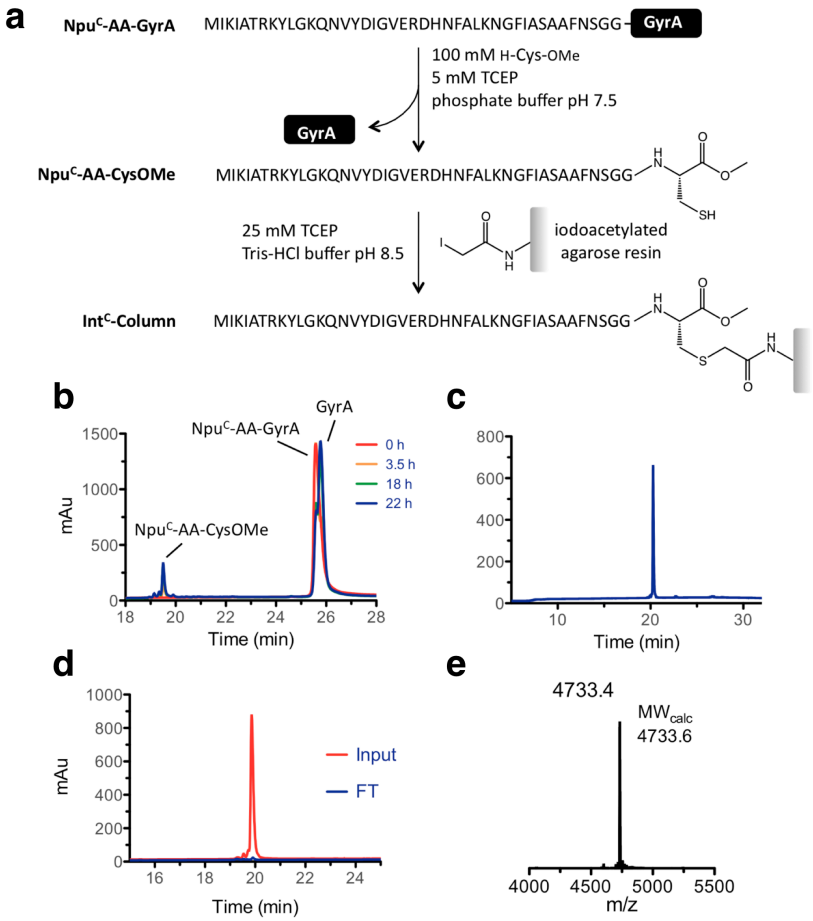


Figure 7.5. Preparation of the Int_C-column. a. Reaction scheme for the production of the Int^C-column. **b.** Time course of the reaction between NpuC-AA-GyrA and H-Cys-OMe monitored by RP-HPLC. **c.** RP-HPLC analysis of the purified NpuC-AA-Cys-OMe peptide. **d.** RP-HPLC analysis of the NpuC-AA-Cys(OMe) solution loaded onto the SulfoLink resin and the column flow-through after immobilization. Less than 3 % of the original input amount was detected in the FT which confirmed that immobilization proceeded with a higher than 97 % yield. All RP-HPLC analyses shown here were performed using a 0-73%B gradient and monitored at 214 nm. **e.** ESI-MS of the purified NpuC-AA-Cys(OMe) peptide before immobilization.

Immobilization was typically performed using a ratio of 0.5 mmols of peptide per mL of SulfoLink resin (final NpuC loading of 0.5 mmols/mL, Figure S5) on a scale of 62.5 to 250 μL of resin beads. Alternatively, Int_C-columns with higher loading (1.8-2.0 mmols/mL) were prepared for the purification of antibody C-terminal α-thioesters, using 2 mmols of peptide per mL of resin.

7.5.4. sEPL experiments

Protocol for proteins expressed in *E. coli*. E. coli BL21(DE3) cells transformed with the desired protein-NpuN plasmid were grown in 1 L of LB containing 100 μg/mL of ampicillin at 37 °C until OD_{600} = 0.6. Protein expression was induced by addition of 0.5 mM IPTG. After harvesting the cells by centrifugation (10,500 rcf, 30 min), the cell pellets were resuspended with 20 mL of high-salt binding buffer (100 mM phosphate, 500 mM NaCl, 1 mM EDTA, 1 mM TCEP, pH 7.2) with Roche complete protease inhibitor cocktail and stored at -80 °C and lysed by sonication. Then the was soluble fraction recovered by centrifugation (17,000 rcf, 10 min). 300 μL of the soluble fraction were loaded onto 62.5 μL of Int_C column and incubated at room temperature for 5 min. After incubation, the flow-through was collected, and the column was washed with 300 μL of high-salt binding buffer, 300 μL of wash buffer (100 mM phosphate, 300 mM NaCl, 1 mM EDTA, 1 mM TCEP, pH 7.2), and 300 μL of binding buffer (100 mM phosphate, 150 mM NaCl, 1 mM EDTA, 1 mM TCEP, pH 7.2). The column was capped and incubated with 150 μL of elution buffer (100 mM phosphate, 150 mM NaCl, 200 mM MESNa, 10 mM TCEP, 1 mM EDTA, pH 7.2) for 18 hours. Flow-through was collected, and the column was washed three times with 150 μL elution buffer. The eluted

thieoster could be directly used for ligation reactions with an N-terminal cystein-containing peptide. For the insoluble histone fragment, the protein was expressed similarly to the other proteins, but after cell lysis, the protein was extracted from the inclusion bodies with 6 M urea then diluted to 2 M urea prior to loading on the Int_C column. Thiolysis from the column required additional incubations with MESNa, as the intein was slower in 2 M urea.

Protocol for sEPL with the monoclonal antibody. HEK293T cells were transiently cotransfected with antimouse-DEC205-LC and antimouse-DEC205-HC-AvaN using lipofectamine 2000 (Invitrogen), according to the manufacturer's instructions. Typical cotransfections were performed in 10 cm plates. After 4 days incubation at 37 °C with 5% CO_2, cell supernatants were harvested and spun down at 2000 rcf for 20 min at 4 °C, filtered through a 0.22 μm filter, and supplemented with Roche complete protease inhibitors. For a typical purification, 50 mL of αDEC205-AvaN transfected cell supernatants were concentrated to a final volume of 5 mL and exchanged into binding buffer. The resulting Ab solution (input) was applied to an Int_C column of 300 μL beads (loading: 1.8 μmol NpuC peptide/mL) and incubated at room temperature for 30 min. Column flow-through was collected, and column washed three times with 3 CV of wash buffer and once with 3 CV of binding buffer. The column was capped and incubated with 3 CV of Ab elution buffer (100 mM phosphate, 150 mM NaCl, 200 mM MES, 1 mM TCEP, 1 mM EDTA, pH 7.2) for 20 h. The column flow-through was collected, and the column washed three times with 3 CV of Ab elution buffer. Elutions containing αDEC205-MES were combined and concentrated down to 20 μM. Ligation was initiated by addition of 1 mM CGK(Fl) peptide and 1 mM TCEP and adjusting pH to 7.5−8.0. The

reaction was incubated in the dark at room temperature for 48 h and monitored by SDS-PAGE imaged using a fluorescence scanner and coomassie staining. Once the reaction was completed, the ligated Ab was diluted to 200−500 µL and dialyzed into 100 mM phosphate, 150 mM NaCl, 1 mM EDTA, 1 mM TCEP, pH 7.2.

7.5.5. Int_C column regeneration

After the desired protein thioesters were eluted from the column, the resin was washed twice with 4 CV of water, and 2 CV of regeneration buffer (40 mM Tris, 10% (v/v) glycerol, 1% (w/v) SDS, pH 7.5). Then the 4 CV of regeneration buffer were added and the column was boiled for at least 7 minutes to reduce the amount of bound NpuN to undetectable levels. After boiling, the column was drained, washed with 2 CV of regeneration buffer, and washed with 4 CV of water three times. The regenerated resin was stored at 4 °C in a pH 7.2 buffer containing 100 mM phosphates, 150 mM NaCl, 1 mM EDTA, 0.05% NaN_3.

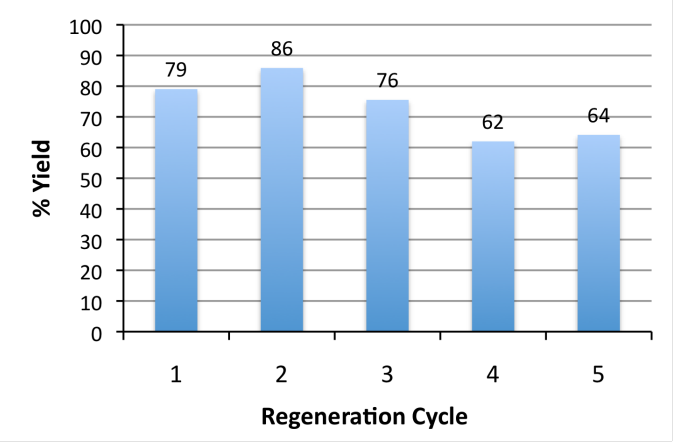


Figure 7.6. Int_C column reuse. The yield of MBP-MES from MBP-Npu_N was quantified over 5 cycles of column regeneration. A constant and known amount of pre-purified MBP-Npu^N was loaded onto the Int^C-column and the eluted MBP-MES was quantified to determine the recovery yield as a function of Int^C-column regeneration. Input and eluted fractions were quantified using the Bradford assay.

The effect of column regeneration on thioester recovery yields was determined using MBP-NpuN. 150 μL of MBP-NpuN at 45 μM (quantified by the Bradford assay) were loaded onto an Int_C-column and purified as above. After purification, the amount of recovered protein was quantified by the Bradford assay and used to calculate recovery yields. The column could be used at least 5 times with only modest loss in loading capacity (Figure 7.6).

7.5.6. Three-piece ligation of PARP1

For all PARP1 ligations, PARP1 fragments were enriched over Ni-NTA resin (see above), and elution fractions from the nickel columns of all three proteins were analyzed on the same Coomassie-stained Bis-Tris gel. Relative concentrations were estimated based on the intensity of product bands in the purest fractions, and these estimates were used to guide the reaction set-up. All reactions were set up in elution buffer (50 mM phosphate, 300 mM NaCl, 250 mM imidazole, no BME, pH 8.0) to achieve a starting material ratio of N : M : C close to 1 eq : 1.5 eq : 1 eq with minimal dilution. After several small-scale tests, a large-scale ligation reaction (6.5 mL) was set up as follows: Starting materials and elution buffer were pre-incubated at 30 °C (without TCEP) for 10 minutes, then the reagents were mixed together with 1 mM TCEP to initiate the reactions. Aliquots of each reaction solution were removed at the desired time points and quenched in a 4x concentrated gel loading dye to afford a final quencher solution with 40 mM Tris (~pH 7.0), 10% (v/v) glycerol, 1% (w/v) SDS, 0.02% (w/v) bromophenol blue, and 2% (v/v) BME.

After 21 hours, the reaction mixture was dialyzed into binding buffer (50 mM

phosphate, 300 mM NaCl, 5 mM imidazole, no BME, pH 8.0) then incubated with 2 mL of Ni-NTA resin for 30 min at 4 °C to trap un-reacted starting materials, intermediates, and spliced intein fragments. The column was washed with 6 mL of binding buffer, then the flow-through and wash fractions were combined and concentrated ~10-fold. The concentrated sample was loaded onto an S200 10/300 column and eluted in PARP1 buffer (20mM Tris HCl, 200mM NaCl, 1mM DTT, pH 8.0). The reaction and purification progress was analyzed by Western blotting against PARP1-N and PARP1-C (antibodies from Santa Cruz Biotech). The purified PARP1 was run along-side commercial PARP1 and BSA standards on a gel, and the gel was Coomassie stained to assess purity and estimate protein concentration. In addition to the clarified three-piece ligation reaction, 0.5 mL of PARP1-C post Ni-NTA column was also purified by size exclusion chromatography in PARP1 buffer. Purification of three-piece-ligated PARP1 from 3 L of bacterial culture yielded roughly 10 μg of pure PARP1. Purified PARP1 was analyzed by LC/MS/MS after trypsinization to confirm the ligated product, and activity assays were carried out as reported previously.[168,169]

7.6. Biophysical and structural characterization

7.6.1. Circular dichroism

Isolated NpuN, isolated NpuC, and an equimolar mixture of the two fragments were dialyzed into CD buffer (25 mM sodium phosphates, 50 mM NaF, 1 mM DTT, pH 7.2). Circular dichroism spectra were measured at 25 °C in a 1 mm pathlength cuvette. The NpuN and complex spectra were measured with 2.5 μM protein and the NpuC spectrum was measured with 12.5 μM protein.

7.6.2. Limited proteolysis

Isolated NpuN, isolated NpuC, and an equimolar mixture of the two fragments were dialyzed into thermolysin buffer (50 mM Tris HCl, 100 mM NaCl, 2 mM $MgSO_4$, 2 mM $CaCl_2$, 1 mM DTT, pH 7.4). Commercially purchased thermolysin powder was dissolved to 0.2 mg/mL in thermolysin buffer and kept on ice until the reaction. The substrate proteins were diluted to 12.5 μM (approximately 0.2 mg/mL of the complex) and incubated at 30 °C for five minutes. The reactions were initiated by addition of thermolysin to each solution to a final ratio of 1:100 (v/v). At various time points, aliquots were removed and quenched by mixing with 1.5 volumes of 0.1% TFA in water to reduce the pH. Quenched time points were analyzed by RP-HPLC and ESI-MS, and the masses were compared to predicted proteolysis products determined using the PeptideCutter tool from the ExPASY server (Figure 7.7).[166]

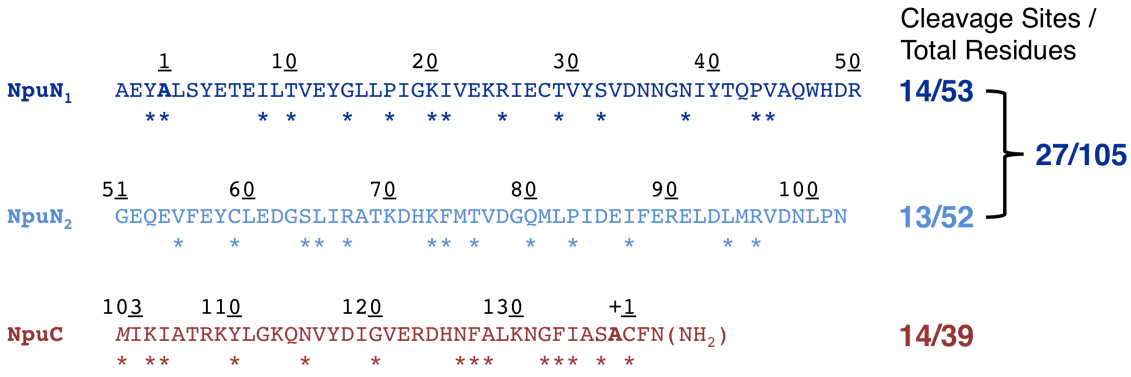


Figure 7.7. Predicted thermolysin cleavage sites on Npu fragments. Sites predicted using the Peptide Cutter online tool are marked with an asterisk.

7.6.3. Equilibrium fluorescence binding measurements

Steady-state tryptophan fluorescence measurements were carried out in splicing assay buffer (100 mM sodium phosphates, 150 mM NaCl, 1 mM EDTA, 1mM DTT, pH 7.2) or a high-salt version with 500 mM NaCl. For titrations of Npu with the Ub and

SUMO exteins, a 50 nM solution of Ub-NpuN was placed in the cuvette, and increasing amounts of NpuC-SUMO were titrated directly into the cuvette, waiting three minutes after NpuC addition before measuring fluorescence spectra. Samples were excited at 295 nm (3 nm bandwith) and emission was recorded between 300 nm and 500 nm (7 nm bandwith). Binding was measured at 30 °C. In these titrations, background fluorescence of NpuC-SUMO was separately measured and subtracted to yield the titration curves. For titrations of Npu with the AEY and CFN exteins, individual solutions of 40 nM AEY-NpuN and a range of NpuC-CFN concentrations were prepared in plastic tubes and pre-incubated for five minutes to 25 °C. Each equilibrated solution was sequentially added to the cuvette and the fluorescence spectrum was immediately measured. This alternate procedure was necessary since free AEY-NpuN adsorbed to the cuvette wall over time. Samples were excited at 290 nm (5 nm bandwith) and emission was recorded between 300 nm and 500 nm (7 nm bandwith). In these titrations, background fluorescence of NpuC-CFN was negligible. With the small exteins, binding was too tight to accurately measure in the 150 mM NaCl buffer, so K_D values were only determined in the 500 mM NaCl buffer. All equilibrium titration curves, measured in triplicate or quadruplicate, were fit to a quadratic binding equation:

$$F_{obs} = F_{NpuN} + \left(F_{Sat} - F_{NpuN}\right)\frac{\left([NpuC] + K_d + [NpuN]\right) - \sqrt{\left([NpuC] + K_d + [NpuN]\right)^2 - 4[NpuN][NpuC]}}{2[NpuN]}$$

ANS fluorescence measurements were carried out in NMR buffer (25 mM sodium phosphates, 100 mM NaCl, 1 mM DTT, pH 6.5) at 30 °C in a quartz cuvette with a 1 cm pathlength. After a five minute incubation at 30 °C, the samples were placed in the cuvette and excited at 370 nm. The emission spectra were recorded between 400 nm and

720 nm. In all measurements, ANS concentration was 4 μM, and fluorescence of the dye was measured alone or in the presence of 4 μM of $NpuN_1$ or $NpuN_2$.

7.6.4. Stopped-flow fluorescence binding measurements

Stopped-flow tryptophan fluorescence measurements were carried out at 25 °C under pseudo-first order conditions using NpuN and NpuC with the short tripeptide exteins. Samples were excited at 290 nm (2 mm slits = ~9 nm bandpass) and emission was recorded above a 320 nm cut-off filter. Measurements were carried out in regular and high salt splicing assay buffer as well as CD buffer (Table 5.2). For each measurement, a 50 nM solution of NpuN was mixed with an equal volume of 250 nM to 1500 nM NpuC using the stopped-flow technique to afford a final NpuN concentration of 25 nM and a 5- to 30-fold excess of NpuC. Trytophan fluorescence was constantly measured for 90 to 300 seconds. Typically, fluorescence curves for three to five sequential mixes from the same batch of proteins were averaged then fit to either a one-phase or two-phase exponential decay.

One-phase exponential:
$$Y = Y_0 + \left(Y_{\max} - Y_0\right)\left(1 - e^{-kt}\right)$$

Two-phase exponential:
$$A_{fast} = \left(Y_{\max} - Y_0\right)F_{fast}$$
$$A_{slow} = \left(Y_{\max} - Y_0\right)\left(1 - F_{fast}\right)$$
$$Y = Y_0 + A_{fast}\left(1 - e^{-k_{fast}t}\right) + A_{slow}\left(1 - e^{-k_{slow}t}\right)$$

where Y is the observed signal, Y_0 is the initial fluorescence, $Y_{\max}$ is the final fluorescence once equilibrium is reached, k is the observed first-order rate constant and t is time. For the two-phase kinetics, A_{fast} and A_{slow} are the fluorescence amplitudes of the fast and slow phases, and F_{fast} and F_{slow} are the fractional amplitudes relative to total fuorescence

change. For the two-phase fits, binding curves for a whole [NpuC] concentration series were globally fit and F_{fast} was constrained to be equal for all curves. For all titrations, the concentration-dependent observed rate constants from multiple titrations were averaged, plotted as a function of NpuC concentration, and fit to a line. The slope of the line was interpreted as the second-order association rate constant k_{on}.

7.6.5. SEC-MALS of Npu fragments and complexes

All SEC experiments, including analyses of standards, were run on an S75 10/300 column at 4 °C in NMR buffer (25 mM sodium phosphates, 100 mM NaCl, 1 mM DTT, pH 6.5) or a high salt version containing 500 mM NaCl when indicated. For all runs, UV absorbance was monitored at 214 nm or 280 nm. For SEC-MALS, a multi-angle light scattering detector and a refractive index detector were on-line immediately following the UV absorbance detector. In a typical experiment, the relevant isolated proteins or mixtures were incubated on ice for 30 minutes at the indicated concentration, then 500 μL of the solution was injected onto an S75 10/300 column and eluted with a flow rate of 0.5 mL/min. For MALS analysis, light scattering and refractive index measurements were averaged across the entire width above half of the peak height to extract the molar mass. For comparison of Stokes radii (R_S) and degree of compaction, the following commercially available, well-characterized standards with the indicated R_S and molecular weight (MW) were used: Vitamin B_{12} (8.5 Å, 1.35 kDa), Aprotinin (13.5 Å, 6.5 kDa), Cytochrome C (17.0 Å, 12.3 kDa), Ribonuclease A (17.5 Å, 13.7 kDa), Myoglobin (19.0 Å, 16.9 kDa), β-lactoglobulin (27.5 Å 35.0 kDa), and Ovalbumin (28.0 Å, 43.0 kDa). A linear relationship between R_S and the inverse of the elution volume was observed

(Figure 7.8a) and used to determine R_S of Npu fragments and complexes (Figure 7.8b). A linear relationship between Log_{10} MW and the inverse of the elution volume for the globular protein standards (thus excluding Vitamin B_{12}) was also observed (Figure 7.8c). This line was used to compare the compactness of Npu fragments and complexes to well-studied globular proteins (Figure 7.8d).

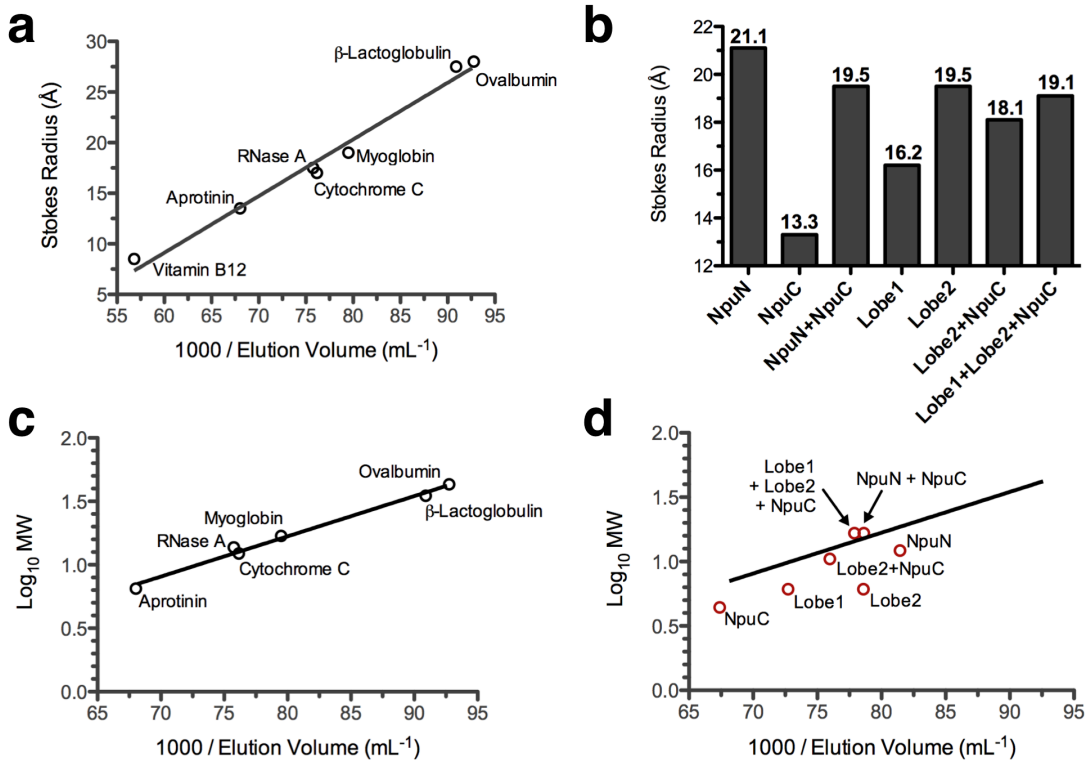


Figure 7.8. SEC standards and the size of Npu fragments/complexes. a. Relationship between experimentally determined elution volumes and published Stokes radii for size-exclusion standards. The solid line is the best fit line to those data. **b.** Stokes radii of proteins and complexes in this study, determined based on the standard line in the previous panel. **c.** Relationship between experimentally determined elution volumes and the calculated molecular weight of globular protein standards. The solid line is the best fit line to those data. **d.** Comparison of the compactness of proteins and complexes in this study relative to the globular standards in the previous panel. Proteins further below the line are less compact than those close to the line.

7.6.6. NMR spectroscopy

General. All NMR data were acquired at 25 °C in NMR Buffer (25 mM sodium

phosphates, 100 mM NaCl, 1 mM DTT, pH 6.5). In all experiments the uniformly labeled protein was concentrated to 250 μM. In experiments where segmentally labeled samples are used the protein concentration was 100 μM. The data was processed with NMRPipe.[170] NMRViewJ was used for relaxation analysis, backbone resonance assignments, and chemical shift index (CSI) calculations.[158]

Resonance Assignments. The backbone assignments of NpuC when alone and in complex with NpuN and $NpuN_2$ were carried out using ^{15}N, ^{13}C labeled fully protonated samples using standard triple resonance experiment pairs: HNCO/HN(CA)CO, HNCA/HN(CO)CA, and HNCACB/CBCA(CO)NH.[171] For the HNCO/HN(CA)CO pair 1024, 42, and 60 complex points recorded with 12.6 ppm, 31 ppm, and 20 ppm sweep widths in ^{1}H, ^{15}N, and ^{13}C dimensions, respectively. For the HNCA/HN(CO)CA pair 1024, 42, and 60 complex points recorded with 12.6 ppm, 31 ppm, and 32 ppm sweep widths in ^{1}H, ^{15}N, and ^{13}C dimensions, respectively. For the HNCACB/CBCA(CO)NH pair 1024, 42, and 60 complex points recorded with 12.6 ppm, 31 ppm, and 58 ppm sweep widths in ^{1}H, ^{15}N, and ^{13}C dimensions, respectively. The backbone and Cβ assignments for $NpuN_2$ when alone and in complex with NpuC were carried out using the same experimental parameters as the NpuC assignments.

Chemical Shift Perturbations. ^{1}H-^{15}N HSQC spectra of samples were recorded at 25 ^{o}C. The chemical shift perturbations ($\Delta\delta_i$) of each assigned residue (i) were calculated using the following equation:

$$\Delta\delta_i = [\ (\Delta\delta_{H,State2} - \Delta\delta_{H,State1})_i^2 + 0.11(\Delta\delta_{N,State2} - \Delta\delta_{N,State1})_i^2\ ]^{1/2}$$

Spin Relaxation Measurements. Relaxation experiments were run at 800 MHz at 25 ^{o}C and the analyses were performed using NMRViewJ. Spin-lattice, spin-spin

relaxation rates (R_1, R_2) and ^{15}N-^{1}H NOE measurements of NpuC when alone and when in complex with NpuN were performed by using standard pulse sequences utilizing two-dimensional ^{1}H-^{15}N correlation.[152] R_1 and R_2 rates were measured using recycle delays of 1.5 s and the following relaxation delays were used:

For the complex: R_1: 10.0, 160, 310, 610 (x2), 760, 910, and 1210 ms

R_2: 0, 16.3, 32.6, 48.9 65.2 (×2), 81.5, and 130.4 ms

For NpuC alone: R_1: 60, 120, 250, 500, 700, 1000, 2000 ms

R_2: 32.6, 65.2, 130.4, 260.8, 391.2, 521.6, 782.4, 1043.2, 1564.8 ms

7.7. Molecular dynamics simulations

The simulations of DnaE intein from *Nostoc punctiforme* (NpuDnaE) were carried out using the molecular dynamics (MD) software package AMBER11.[172] The first alternative NMR solution structure within PDB 2KEQ was selected.[39] The selected fused structure was cut between Asn102 and Ile103 to mimic the split form of the intein. The sequence was modified *in silico* using UCSF Chimera[155] to generate constructs of interest: native Npu with the canonical N-extein ($A_{-3}E_{-2}Y_{-1}$) and C-extein ($C_{+1}F_{+2}N_{+3}$) sequences, a C-extein variant ($C_{+1}A_{+2}N_{+3}$), or a D124Y mutation combined with the same C-extein mutation. The C-terminal ends of the C-exteins were charge capped using a special C-terminal residue, NHE, defined in the AMBER ff99sb force field.[173] As discussed in Chapter 3, identical all-atom molecular dynamics (MD) simulations were run using standard techniques in an explicit water box charge neutralized with sodium ions. The simulation time step was 1 fs and snap shots were saved at every 5 ps. The distances and dihedral angles were extracted from these snap shots over the MD trajectories.[174-177]

7.8. Bioinformatic analyses

7.8.1. Charge segregation in split and contiguous inteins

All mini-intein sequences (116 total) were retrieved from the InBase based on the criteria that they were explicitly annotated as having no Block C, D, E, or H motifs (the homing endonuclease motifs).[10] Sequences longer than 200 residues (14 sequences) were removed to avoid alignment problems due to large insertions. The remaining sequences were aligned with a Hidden Markov Model (HMM) from the SUPERFAMILY database (model # 0036550)[178] using HMMER 3.0.[179,180] Several mis-aligned sequences (25%) were removed from the resulting alignment, and the remaining sequence alignment was manually optimized based on known intein sequence motifs.[10] This optimized alignment was used to build a new HMM, which in turn was used to align all 116 mini-intein sequences extracted from the InBase. The resulting alignment was manually optimized, and this optimized alignment (24 split inteins and 76 intact inteins) was used for all bioinformatic analyses.

To analyze the isoelectric points (pI) of N- and C-intein sequences, the sequence alignment was split into two alignments: one alignment contained all positions that aligned with the N-terminal methionine of the C-inteins through the C-terminal asparagine; the second alignment contained all sequence positions before the C-intein region. The isoelectric point of each sequence fragment was estimated using the pI/MW tool on the ExPASY server.[166] To generate sequence logos, these split alignments were further modified, removing any positions that did not align with the Npu_{WT} intein. The modified alignments were separated based on naturally split and intact inteins and represented as sequence logos using WebLogo.[181]

7.8.2. Charge-hydrophobicity analysis

For the charge-hydrophobicity analysis of Npu fragments and complexes, mean net charge and mean hydrophobicity were calculated as previously described.[3] Mean net charge (R) is the absolute value of the difference between the number of Arg/Lys residues and Asp/Glu residues in a sequence divided by the total number of amino acids. Hydrophobicity is calculated on a normalized Kyte-Doolittle scale[182] from 0 (most polar) to 1 (most hydrophobic), and mean hydrophobicity (H) is the average normalized hydrophobicity over the whole sequence. The solid line delineating disordered and folded proteins is empirically defined by Uversky et al. as $R = 2.785 \cdot H - 1.151$ (Figure 5.5a).[3]

7.9. Analytical data for purified proteins and peptides

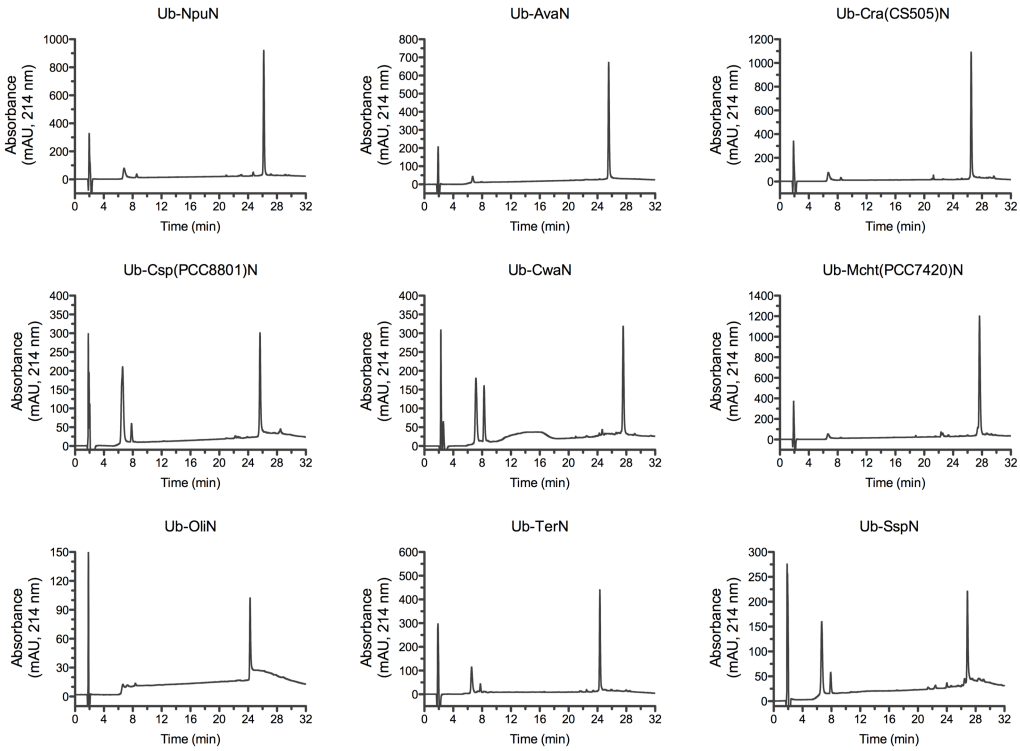


Figure 7.9. RP-HPLC analysis of the pure Ub-IntN fusions in Chapter 2. Proteins were run on an analytical C18 column over a 0-73% B gradient in 30 minutes. Peaks between 6 and 9 minutes are from DTT in the sample buffer.

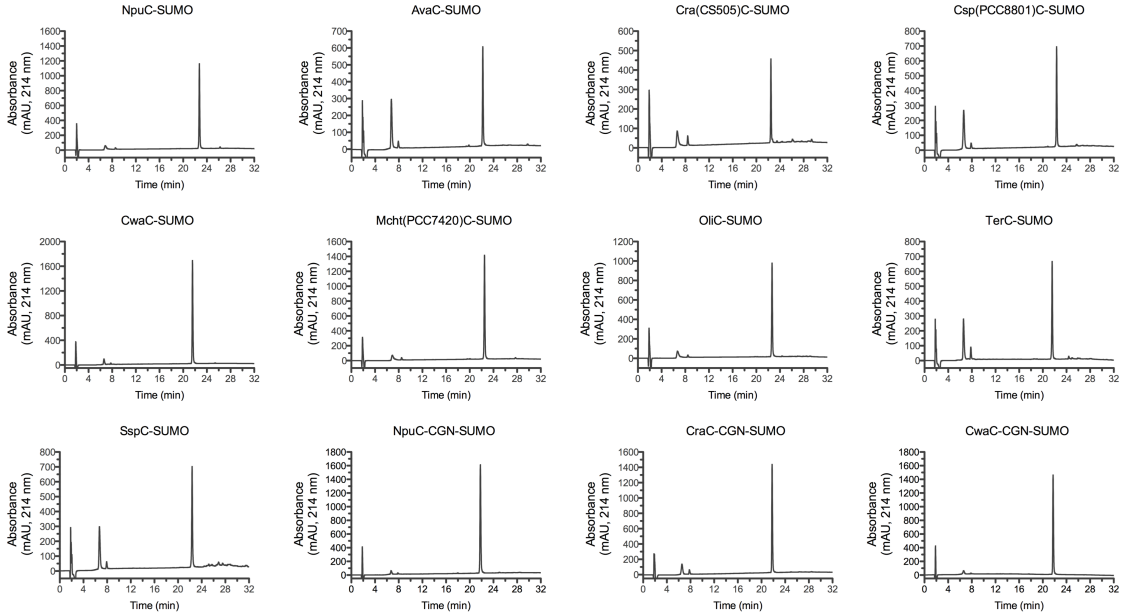


Figure 7.10. RP-HPLC analysis of the pure IntC-SUMO fusions in Chapter 2. Proteins were run on an analytical C18 column over a 0-73% B gradient in 30 minutes. Peaks between 6 and 9 minutes are from DTT in the sample buffer.

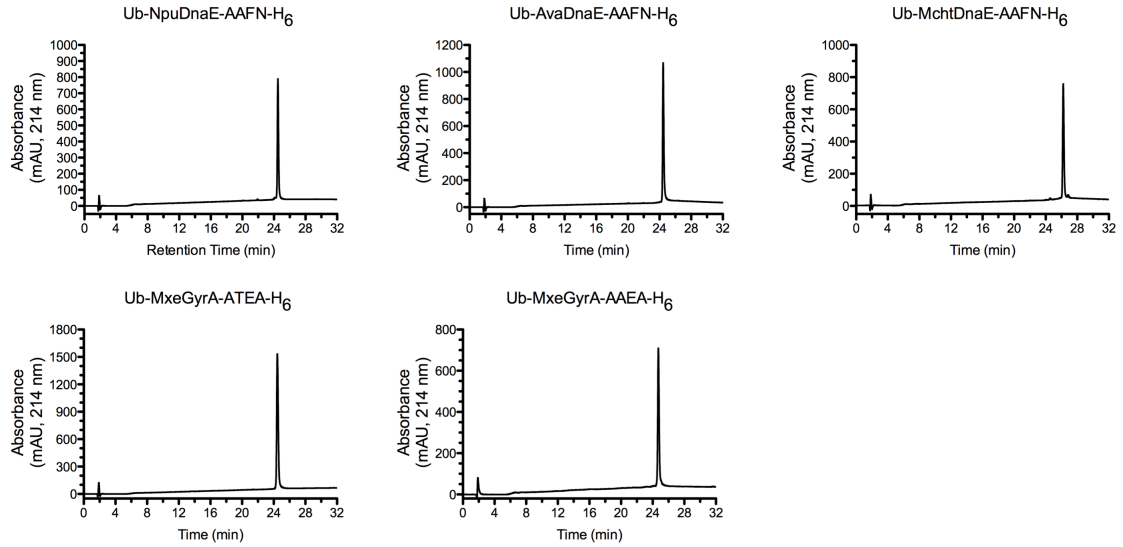


Figure 7.11. RP-HPLC analysis of Ub-contiguous intein fusions in Chapter 2. All proteins were diluted 20-fold in H_2O with 0.1% TFA and incubated at room temperature for a minimum of 2 hours then injected onto an analytical C18 column. The proteins were eluted over a 30 minute 0-73% buffer B gradient.

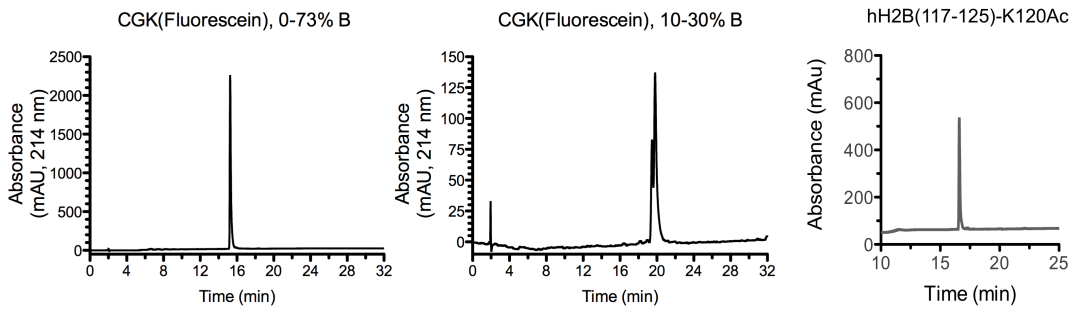


Figure 7.12. RP-HPLC analysis of peptides used in EPL/sEPL studies in Chapter 2. Proteins were run on an analytical C18 column over a 0-73% B or 10-30% B gradient in 30 minutes.

Table 7.1. Masses of purified proteins/peptides from Chapter 2.[a]

Protein/Peptide	Expected Mass (Da)	Observed Mass (Da)
Ub-NpuN	22074.97	22074.7
Ub-AvaN	21829.76	21829.6
Ub-Cra(CS505)N	23734.97	23734.7
Ub-Csp(PCC8801)N	21734.69	21734.6
Ub-CwaN	22537.41	22537.3
Ub-Mcht(PCC7420)N	22225.02	22224.9
Ub-OliN	22949.89	22949.8
Ub-TerN	22020.98	22020.8
Ub-SspN	24186.24	24186.8
NpuC-SUMO	17655.87	17655.7
AvaC-SUMO	17553.77	17553.6
Cra(CS505)C-SUMO	17535.80	17535.8
SspC-SUMO	17435.69	17435.4
Csp(PCC8801)C-SUMO	17486.64	17486.7
CwaC-SUMO	17422.61	17422.5
Mcht(PCC7420)C-SUMO	17521.66	17521.4
OliC-SUMO	17521.66	17520.8
TerC-SUMO	17489.73	17489.7
NpuC-CGN-SUMO	17565.74	17565.8
Cra(CS505)C-CGN-SUMO	17445.68	17445.7
CwaC-CGN-SUMO	17332.48	17332.5
Ub-NpuDnaE-AAFN-H_6	25491.98	25492.2
Ub-AvaDnaE-AAFN-H_6	25144.67	25144.8
Ub-MchtDNaE-AAFN-H_6	25507.82	25508.1
Ub-MxeGyrA-ATEA-H_6	30941.03	30940.9
Ub-MxeGyrA-AAEA-H_6	30911.00	30910.9
SH3-AvaDnaE-AAFN-H_6	23000.07	23000.1
SH3-MchtDnaE-AAFN-H_6	23363.22	23363.1
SH2-AvaDnaE-AAFN-H_6	27487.92	27488.3
eGFP-AvaDnaE-AAFN-H_6	43501.28	43501.9
$PARP1_C$-AvaDnaE-AAFN-H_6	56681.87	56681.1
CGK(Fl) [b]	663.20	664.2
hH2B(117-125)-K120Ac [b]	1041.5	1042.5

[a] Note that the calculated and observed masses are the average masses rather than the monoisotopic masses.

[b] The observed mass is the singly charged $[M+H]^+$ species.

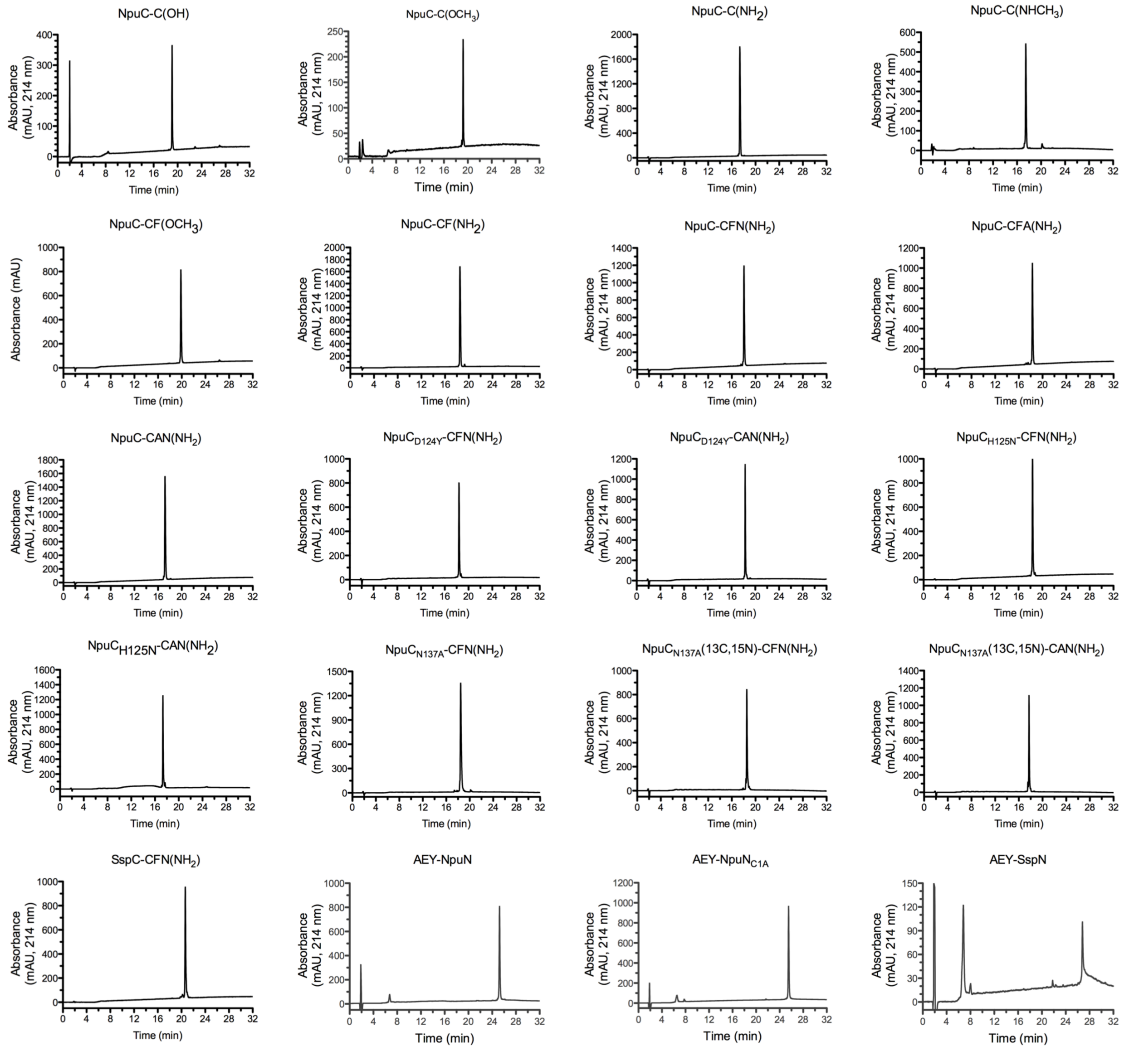


Figure 7.13. RP-HPLC analysis of the pure proteins in Chapter 3. Proteins were run on an analytical C18 column over a 0-73% B gradient in 30 minutes. Peaks between 6 and 9 minutes are from DTT in the sample buffer for the in the N-inteins.

Table 7.2. Masses of purified proteins/peptides from Chapter 3.

Protein/Peptide	Expected Mass (Da)	Observed Mass (Da)
AEY-NpuN [a]	12219.8	12219.5
AEY-NpuN(C1A) [a]	12187.7	12187.4
AEY-SspN [a]	14331.1	14330.6
NpuC-C(OH)	4226.2	4226.2
NpuC-C(OCH_3)	4240.2	4240.2
NpuC-C(NH_2)	4225.2	4225.2
NpuC-C($NHCH_3$)	4239.2	4239.2
NpuC-CF(OCH_3)	4387.3	4387.3
NpuC-CF(NH_2)	4372.3	4372.3
NpuC-CFN(NH_2)	4486.3	4486.4
NpuC-CFA(NH_2)	4443.3	4443.4
NpuC-CAN(NH2)	4410.3	4410.4
NpuC(D124Y)-CFN(NH_2)	4534.4	4534.4
NpuC(D124Y)-CAN(NH_2)	4458.3	4458.3
NpuC(H125N)-CFN(NH_2)	4463.3	4463.3
NpuC(H125N)-CAN(NH_2)	4387.3	4387.3
NpuC(N137A)-CFN(NH_2)	4443.3	4443.4
NpuC(N137A)($^{13}C,^{15}N$)-CFN(NH_2) [b]	4679.8	4675.8
NpuC(N137A)($^{13}C,^{15}N$)-CAN(NH_2) [b]	4603.7	4599.7
SspC-CFN(NH2) [c]	4135.3	4135.3

[a] The expected and observed masses for the N-intein constructs are the average mass. The expected and observed masses of C-intein constructs are the monoisotopic masses.

[b] Based on the difference in expected and observed masses, the ^{13}C and ^{15}N isotopic enrichment for these samples is >98%.

[c] $SspC_{WT}$ was isolated without its N-terminal methionine due to significant *in vivo* processing during expression in *E. coli*.

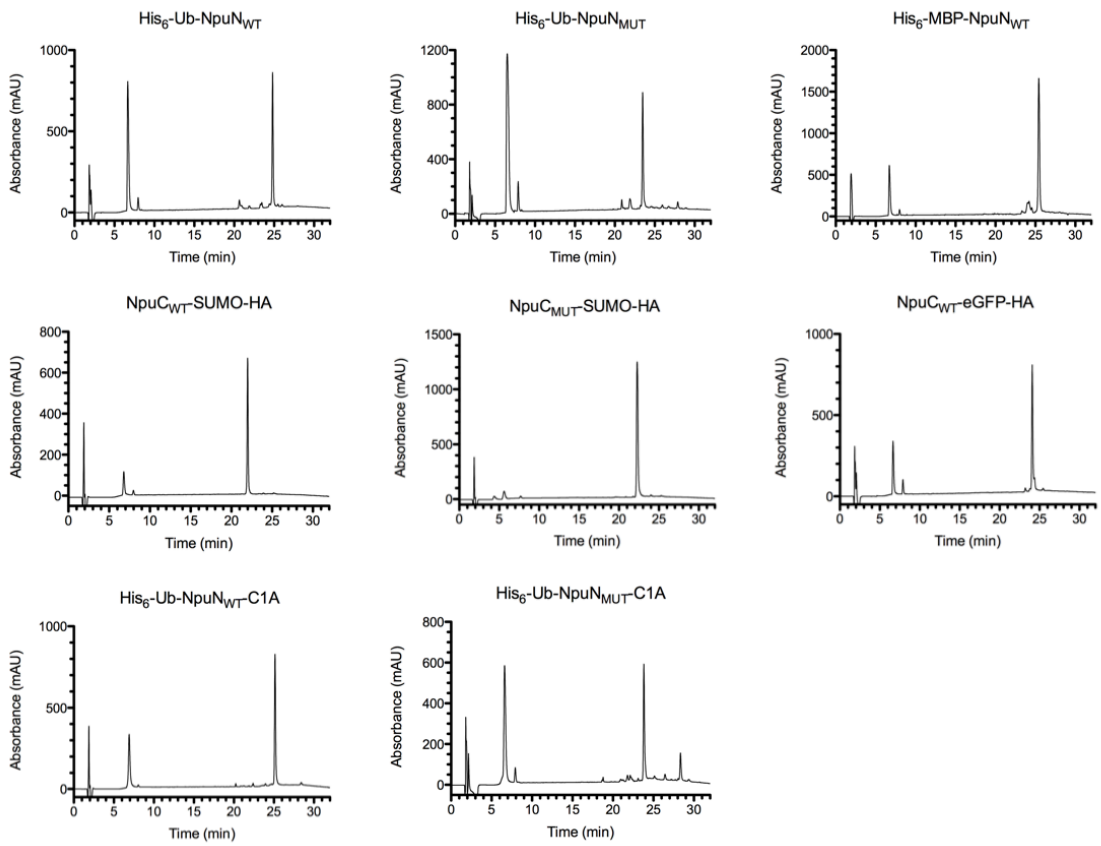


Figure 7.14. RP-HPLC analysis of the pure IntN and IntC fusions in Chapter 4. Proteins were run on an analytical C18 column over a 0-73% B gradient in 30 minutes. Peaks between 6 and 9 minutes are from DTT in the sample buffer.

Table 7.3. Masses of purified proteins/peptides from Chapter 4.[a]

Protein/Peptide	Expected Mass (Da)	Observed Mass (Da)
His_6-Ub-$NpuN_{WT}$ [b]	22303.1	22301.9
His_6-Ub-$NpuN_{MUT}$ [b]	22338.6	22337.4
His_6-Ub-$NpuN_{WT}$-C1A [b]	22271.1	22272.3
His_6-Ub-$NpuN_{MUT}$C1A [b]	22306.6	22307.0
$NpuC_{WT}$-SUMO-HA	17655.8	17656.4
$NpuC_{MUT}$-SUMO-HA	17631.6	17635.3
His_6-MBP-$NpuN_{WT}$ [b]	54509.6	54526.0
$NpuC_{WT}$-eGFP-HA [c]	33317.6	33321.8
SH3-$NpuN_{MUT}$-His_6	19493.5	19495.7
His_6-$NpuC_{MUT}$-GB1-$NpuN_{WT}$-His_6	24736.5	24739.2
His_6-TEV-$NpuC_{WT}$-eGFP-HA	35224.9	35220.2

[a] All purified proteins in this chapter were analyzed on the Sciex-API-100 mass spectrometer, which was significantly older than the Bruker MicrOTOF-Q instrument used in other chapters, and thus the error was greater for these proteins.

[b] These proteins differ from the Protein-IntN constructs in Table 7.1 in that they have a Gly_4 linker between the extein domain and the canonical KFAEY N-extein sequence before IntN

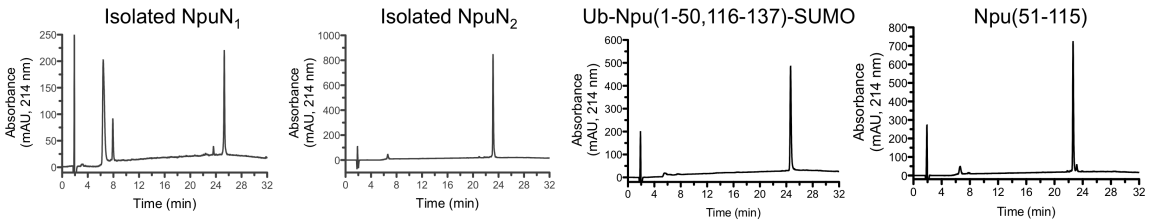


Figure 7.15. RP-HPLC analysis of lobes and intein permutations in Chapter 5. Proteins were run on an analytical C18 column over a 0-73% B gradient in 30 minutes. Peaks between 6 and 9 minutes are from DTT in the sample buffer.

Table 7.4. Masses of purified proteins/peptides from Chapter 5.

Protein/Peptide	Expected Mass (Da)	Observed Mass (Da)
AEY-NpuN(C1A) ($^{13}C,^{15}N$) [a]	12861.7	12853.2
AEY-NpuN(C1A) ($^{13}C,^{15}N$ *on NpuN$_1$*) [a,b]	12852.8	12578.3
AEY-NpuN(C1A) ($^{13}C,^{15}N$ *on NpuN$_2$*) [a,b]	12466.6	12462.7
AEY-NpuN(C1A-50)	6100.1	6100.1
NpuN(51-102)	6097.9	6097.9
NpuN(51-102) ($^{13}C,^{15}N$) [b]	6433.6	6427.7
Npu(51-115)	7611.9	7611.8
Npu(51-115) ($^{13}C,^{15}N$) [b]	8037.7	8028.7
SUMO-AEY-NpuN(1-50) [a]	19402.6	19402.6
Ub-Npu(1-50,116-137)-SUMO [a]	32171.1	32171.2

[a] The expected and observed masses for the N-intein constructs are the average mass. The expected and observed masses of C-intein constructs are the monoisotopic masses.

[b] Based on the difference in expected and observed masses, the ^{13}C and ^{15}N isotopic enrichment for these samples is >98%.

References

1 Anfinsen, C. B., Principles that govern the folding of protein chains. *Science* **181**, 223-230 (1973).

2 Hartl, F. U. and Hayer-Hartl, M., Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.* **16**, 574-581 (2009).

3 Uversky, V. N., Gillespie, J. R., and Fink, A. L., Why are "natively unfolded" proteins unstructured under physiologic conditions? *Protein Struct. Funct. Genet.* **41**, 415-427 (2000).

4 Walsh, C. T., Garneau-Tsodikova, S., and Gatto, G. J., Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem. Int. Ed.* **44**, 7342-7372 (2005).

5 Waugh, D. S., An overview of enzymatic reagents for the removal of affinity tags. *Protein Expr. Purif.* **80**, 283-293 (2011).

6 Schlosser, A., Vanselow, J. T., and Kramer, A., Mapping of phosphorylation sites by a multi-protease approach with specific phosphopeptide enrichment and NanoLC-MS/MS analysis. *Anal. Chem.* **77**, 5243-5250 (2005).

7 Rao, M. B., Tanksale, A. M., Ghatge, M. S., and Deshpande, V. V., Molecular and biotechnological aspects of microbial proteases. *Microbiol. Mol. Biol. Rev.* **62**, 597-635 (1998).

8 Vila-Perelló, M. and Muir, T. W., Biological applications of protein splicing. *Cell* **143**, 191-200 (2010).

9 Kane, P. M. et al., Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* **250**, 651-657 (1990).

10 Perler, F. B., InBase: The intein database. *Nucleic Acids Res.* **30**, 383-384 (2002).

11 Pietrokovski, S., Intein spread and extinction in evolution. *Trends Genet.* **17**, 465-472 (2001).

12 Liu, X.-Q. and Yang, J., Split dnaE genes encoding multiple novel inteins in Trichodesmium erythraeum. *J. Biol. Chem.* **278**, 26315-26318 (2003).

13 Letunic, I. and Bork, P., Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128 (2007).

14 Dalgaard, J. Z., Moser, M. J., Hughey, R., and Mian, I. S., Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J. Comput. Biol.* **4**, 193-214 (1997).

15 Frischkorn, K. et al., Investigation of mycobacterial recA function: protein introns in the RecA of pathogenic mycobacteria do not affect competency for homologous recombination. *Mol. Microbiol.* **29**, 1203-1214 (1998).

16 Papavinasasundaram, K. G., Colston, M. J., and Davis, E. O., Construction and complementation of a recA deletion mutant of Mycobacterium smegmatis reveals

that the intein in Mycobacterium tuberculosis recA does not affect RecA function. *Mol. Microbiol.* **30**, 525-534 (1998).

17 Dassa, B., London, N., Stoddard, B. L., Schueler-Furman, O., and Pietrokovski, S., Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. *Nucleic Acids Res.* **37**, 2560-2573 (2009).

18 Ogata, H., Raoult, D., and Claverie, J.-M., A new example of viral intein in Mimivirus. *Virol. J.* **2**, 8 (2005).

19 Caspi, J., Amitai, G., Belenkiy, O., and Pietrokovski, S., Distribution of split DnaE inteins in cyanobacteria. *Mol. Microbiol.* **50**, 1569-1577 (2003).

20 Choi, J. J. et al., Protein trans-splicing and characterization of a split family B-type DNA polymerase from the hyperthermophilic archaeal parasite Nanoarchaeum equitans. *J. Mol. Biol.* **356**, 1093-1106 (2006).

21 Perler, F. B., A natural example of protein trans-splicing. *Trends Biochem. Sci.* **24**, 209-211 (1999).

22 Volkmann, G. and Mootz, H. D., Recent progress in intein research: from mechanism to directed evolution and applications. *Cell. Mol. Life Sci.* (2012).

23 Southworth, M. W., Benner, J., and Perler, F. B., An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile. *EMBO J.* **19**, 5019-5026 (2000).

24 Tori, K. et al., Splicing of the mycobacteriophage Bethlehem DnaB intein: identification of a new mechanistic class of inteins that contain an obligate block F nucleophile. *J. Biol. Chem.* **285**, 2515-2526 (2010).

25 Romanelli, A., Shekhtman, A., Cowburn, D., and Muir, T. W., Semisynthesis of a segmental isotopically labeled protein splicing precursor: NMR evidence for an unusual peptide bond at the N-extein-intein junction. *Proc. Natl. Acad. Sci. USA* **101**, 6397-6402 (2004).

26 Dearden, A. K. et al., A conserved threonine spring-loads precursor for intein splicing. *Protein Sci.* **22**, 557-563 (2013).

27 Johnson, M. A. et al., NMR structure of a KlbA intein precursor from Methanococcus jannaschii. *Protein Sci.* **16**, 1316-1328 (2007).

28 Amitai, G., Dassa, B., and Pietrokovski, S., Protein splicing of inteins with atypical glutamine and aspartate C-terminal residues. *J. Biol. Chem.* **279**, 3121 (2004).

29 Chen, L., Benner, J., and Perler, F. B., Protein splicing in the absence of an intein penultimate histidine. *J. Biol. Chem.* **275**, 20431-20435 (2000).

30 Frutos, S., Goger, M., Giovani, B., Cowburn, D., and Muir, T. W., Branched intermediate formation stimulates peptide bond cleavage in protein splicing. *Nat. Chem. Biol.* **6**, 527 (2010).

31 Chong, S. et al., Protein splicing involving the Saccharomyces cerevisiae VMA intein. The steps in the splicing pathway, side reactions leading to protein

cleavage, and establishment of an in vitro splicing system. *J. Biol. Chem.* **271**, 22159-22168 (1996).

32 Chong, S., Williams, K. S., Wotkowicz, C., and Xu, M. Q., Modulation of protein splicing of the Saccharomyces cerevisiae vacuolar membrane ATPase intein. *J. Biol. Chem.* **273**, 10567-10577 (1998).

33 Muir, T. W., Sondhi, D., and Cole, P. A., Expressed protein ligation: a general method for protein engineering. *Proc. Natl. Acad. Sci. USA* **95**, 6705-6710 (1998).

34 Wood, D. W., Wu, W., Belfort, G., Derbyshire, V., and Belfort, M., A genetic system yields self-cleaving inteins for bioseparations. *Nat. Biotechnol.* **17**, 889-892 (1999).

35 Hall, T. M. et al., Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins. *Cell* **91**, 85-97 (1997).

36 Matsumura, H. et al., Crystal structure of intein homing endonuclease II encoded in DNA polymerase gene from hyperthermophilic archaeon Thermococcus kodakaraensis strain KOD1. *Proteins* **63**, 711-715 (2006).

37 Klabunde, T., Sharma, S., Telenti, A., Jacobs, W. R., and Sacchettini, J. C., Crystal structure of GyrA intein from Mycobacterium xenopi reveals structural basis of protein splicing. *Nat. Struct. Biol.* **5**, 31-36 (1998).

38 Sun, P. et al., Crystal structures of an intein from the split dnaE gene of Synechocystis sp. PCC6803 reveal the catalytic model without the penultimate histidine and the mechanism of zinc ion inhibition of protein splicing. *J. Mol. Biol.* **353**, 1093-1105 (2005).

39 Oeemig, J. S., Aranko, A. S., Djupsjöbacka, J., Heinämäki, K., and Iwaï, H., Solution structure of DnaE intein from Nostoc punctiforme: Structural basis for the design of a new split intein suitable for site-specific chemical modification. *FEBS Lett.* **583**, 1451-1456 (2009).

40 Aranko, A. S., Oeemig, J. S., and Iwaï, H., Structural basis for protein trans-splicing by a bacterial intein-like domain--protein ligation without nucleophilic side chains. *FEBS J.* **280**, 3256-3269 (2013).

41 Amitai, G., Belenkiy, O., Dassa, B., Shainskaya, A., and Pietrokovski, S., Distribution and function of new bacterial intein-like protein domains. *Mol. Microbiol.* **47**, 61-73 (2003).

42 Dassa, B., Haviv, H., Amitai, G., and Pietrokovski, S., Protein splicing and auto-cleavage of bacterial intein-like domains lacking a C'-flanking nucleophilic residue. *J. Biol. Chem.* **279**, 32001-32007 (2004).

43 Lee, J. J. et al., Autoproteolysis in hedgehog protein biogenesis. *Science* **266**, 1528-1537 (1994).

44 Koonin, E. V., A protein splice-junction motif in hedgehog family proteins. *Trends Biochem. Sci.* **20**, 141-142 (1995).

45 Mann, R. K. and Beachy, P. A., Cholesterol modification of proteins. *Biochim. Biophys. Acta* **1529**, 188-202 (2000).

46 Snell, E. A. et al., An unusual choanoflagellate protein released by Hedgehog autocatalytic processing. *Proc. Biol. Sci.* **273**, 401-407 (2006).

47 King, N. and Carroll, S. B., A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. *Proc. Natl. Acad. Sci. USA* **98**, 15032-15037 (2001).

48 Callahan, B. P., Topilina, N. I., Stanger, M. J., Van Roey, P., and Belfort, M., Structure of catalytically competent intein caught in a redox trap with functional and evolutionary implications. *Nat. Struct. Mol. Biol.* (2011).

49 Aranko, A. S., Oeemig, J. S., Kajander, T., and Iwaï, H., Intermolecular domain swapping induces intein-mediated protein alternative splicing. *Nat. Chem. Biol.* **9**, 616-622 (2013).

50 Shi, J. and Muir, T. W., Development of a tandem protein trans-splicing system based on native and engineered split inteins. *J. Am. Chem. Soc.* **127**, 6198-6206 (2005).

51 Davis, E. O., Jenner, P. J., Brooks, P. C., Colston, M. J., and Sedgwick, S. G., Protein splicing in the maturation of M. tuberculosis recA protein: a mechanism for tolerating a novel class of intervening sequence. *Cell* **71**, 201-210 (1992).

52 Xu, M. Q., Southworth, M. W., Mersha, F. B., Hornstra, L. J., and Perler, F. B., In vitro protein splicing of purified precursor and the identification of a branched intermediate. *Cell* **75**, 1371-1377 (1993).

53 Cooper, A. A., Chen, Y. J., Lindorfer, M. A., and Stevens, T. H., Protein splicing of the yeast TFP1 intervening protein sequence: a model for self-excision. *EMBO J.* **12**, 2575-2583 (1993).

54 Southworth, M., Amaya, K., Evans, T., Xu, M., and Perler, F., Purification of proteins fused to either the amino or carboxy terminus of the Mycobacterium xenopi gyrase A intein. *BioTechniques* **27**, 110-120 (1999).

55 Dawson, P. E., Muir, T. W., Clark-Lewis, I., and Kent, S. B., Synthesis of proteins by native chemical ligation. *Science* **266**, 776-779 (1994).

56 Muir, T. W., Semisynthesis of proteins by expressed protein ligation. *Annu. Rev. Biochem.* **72**, 249-289 (2003).

57 Dawson, P., Churchill, M., Ghadiri, M., and Kent, S., Modulation of reactivity in native chemical ligation through the use of thiol additives. *J. Am. Chem. Soc.* **119**, 4325-4329 (1997).

58 Mootz, H. D., Split inteins as versatile tools for protein semisynthesis. *ChemBioChem* **10**, 2579-2589 (2009).

59 Appleby, J. H., Zhou, K., Volkmann, G., and Liu, X.-Q., Novel split intein for trans-splicing synthetic peptide onto C terminus of protein. *J. Biol. Chem.* **284**, 6194-6199 (2009).

60 Aranko, A. S., Züger, S., Buchinger, E., and Iwaï, H., In vivo and in vitro protein ligation by naturally occurring and engineered split DnaE inteins. *PLoS ONE* **4**, e5185 (2009).

61 Ludwig, C., Pfeiff, M., Linne, U., and Mootz, H. D., Ligation of a synthetic peptide to the N terminus of a recombinant protein using semisynthetic protein trans-splicing. *Angew. Chem. Int. Ed.* **45**, 5218-5221 (2006).

62 Ludwig, C., Schwarzer, D., and Mootz, H. D., Interaction studies and alanine scanning analysis of a semi-synthetic split intein reveal thiazoline ring formation from an intermediate of the protein splicing reaction. *J. Biol. Chem.* **283**, 25264-25272 (2008).

63 Jansson, M. et al., High-level production of uniformly ^{15}N- and ^{13}C-enriched fusion proteins in Escherichia coli. *J. Biomol. NMR* **7**, 131-141 (1996).

64 Xu, R., Ayers, B., Cowburn, D., and Muir, T. W., Chemical ligation of folded recombinant proteins: segmental isotopic labeling of domains for NMR studies. *Proc. Natl. Acad. Sci. USA* **96**, 388-393 (1999).

65 Yamazaki, T. et al., Segmental isotope labeling for protein NMR using peptide splicing. *J. Am. Chem. Soc.* **120**, 5591-5592 (1998).

66 Liu, D., Xu, R., and Cowburn, D., Segmental isotopic labeling of proteins for nuclear magnetic resonance. *Meth. Enzymol.* **462**, 151-175 (2009).

67 Volkmann, G. and Iwaï, H., Protein trans-splicing and its use in structural biology: opportunities and limitations. *Mol. Biosyst.* **6**, 2110-2121 (2010).

68 Züger, S. and Iwai, H., Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. *Nat. Biotechnol.* **23**, 736-740 (2005).

69 Camarero, J. and Muir, T., Biosynthesis of a head-to-tail cyclized protein with improved biological activity. *J. Am. Chem. Soc.* **121**, 5597-5598 (1999).

70 Camarero, J., Pavel, J., and Muir, T. W., Chemical synthesis of a circular protein domain: evidence for folding-assisted cyclization. *Angew. Chem. Int. Ed.* **37**, 347-349 (1998).

71 Scott, C. P., Abel-Santos, E., Wall, M., Wahnon, D. C., and Benkovic, S. J., Production of cyclic peptides and proteins in vivo. *Proc. Natl. Acad. Sci. USA* **96**, 13638-13643 (1999).

72 Jeffries, C. M. et al., Stabilization of a binary protein complex by intein-mediated cyclization. *Protein Sci.* **15**, 2612-2618 (2006).

73 Naumann, T. A., Tavassoli, A., and Benkovic, S. J., Genetic selection of cyclic peptide Dam methyltransferase inhibitors. *ChemBioChem* **9**, 194-197 (2008).

74 Young, T. S. et al., Evolution of cyclic peptide protease inhibitors. *Proc. Natl. Acad. Sci. USA* **108**, 11052-11056 (2011).

75 Tavassoli, A. et al., Inhibition of HIV budding by a genetically selected cyclic peptide targeting the Gag-TSG101 interaction. *ACS Chem. Biol.* **3**, 757-764 (2008).

76 Kritzer, J. A. et al., Rapid selection of cyclic peptides that reduce alpha-synuclein toxicity in yeast and animal models. *Nat. Chem. Biol.* **5**, 655-663 (2009).

77 Skretas, G. and Wood, D. W., Regulation of protein activity with small-molecule-controlled inteins. *Protein Sci.* **14**, 523-532 (2005).

78 Buskirk, A. R., Ong, Y.-C., Gartner, Z. J., and Liu, D. R., Directed evolution of ligand dependence: small-molecule-activated protein splicing. *Proc. Natl. Acad. Sci. USA* **101**, 10505-10510 (2004).

79 Cook, S. N. et al., Photochemically initiated protein splicing. *Angew. Chem. Int. Ed.* **34**, 1629-1630 (1995).

80 Vila-Perelló, M., Hori, Y., Ribó, M., and Muir, T. W., Activation of protein splicing by protease- or light-triggered O to N acyl migration. *Angew. Chem. Int. Ed.* **47**, 7764-7767 (2008).

81 Mootz, H. D. and Muir, T. W., Protein splicing triggered by a small molecule. *J. Am. Chem. Soc.* **124**, 9044-9045 (2002).

82 Mootz, H. D., Blum, E. S., Tyszkiewicz, A. B., and Muir, T. W., Conditional protein splicing: a new tool to control protein structure and function in vitro and in vivo. *J. Am. Chem. Soc.* **125**, 10561-10569 (2003).

83 Tyszkiewicz, A. B. and Muir, T. W., Activation of protein splicing with light in yeast. *Nat. Methods* **5**, 303-305 (2008).

84 Prescher, J. A. and Bertozzi, C. R., Chemistry in living systems. *Nat. Chem. Biol.* **1**, 13-21 (2005).

85 Giriat, I. and Muir, T. W., Protein semi-synthesis in living cells. *J. Am. Chem. Soc.* **125**, 7180-7181 (2003).

86 Borra, R., Dong, D., Elnagar, A. Y., Woldemariam, G. A., and Camarero, J. A., In-cell fluorescence activation and labeling of proteins mediated by FRET-quenched split inteins. *J. Am. Chem. Soc.* **134**, 6344-6353 (2012).

87 Li, J., Sun, W., Wang, B., Xiao, X., and Liu, X.-Q., Protein trans-splicing as a means for viral vector-mediated in vivo gene therapy. *Hum. Gene Ther.* **19**, 958 (2008).

88 Martin, D. D., Xu, M. Q., and Evans, T. C., Characterization of a naturally occurring trans-splicing intein from Synechocystis sp. PCC6803. *Biochemistry* **40**, 1393-1402 (2001).

89 Zettler, J., Schütz, V., and Mootz, H. D., The naturally split Npu DnaE intein exhibits an extraordinarily high rate in the protein trans-splicing reaction. *FEBS Lett.* **583**, 909-914 (2009).

90 Iwai, H., Züger, S., Jin, J., and Tam, P.-H., Highly efficient protein trans-splicing by a naturally split DnaE intein from Nostoc punctiforme. *FEBS Lett.* **580**, 1853-1858 (2006).

91 Amitai, G., Callahan, B. P., Stanger, M. J., Belfort, G., and Belfort, M., Modulation of intein activity by its neighboring extein substrates. *Proc. Natl. Acad. Sci. USA* **106**, 11005-11010 (2009).

92 Lockless, S. W. and Muir, T. W., Traceless protein splicing utilizing evolved split inteins. *Proc. Natl. Acad. Sci. USA* **106**, 10999-11004 (2009).

93 Muona, M., Aranko, A. S., Raulinaitis, V., and Iwaï, H., Segmental isotopic labeling of multi-domain and fusion proteins by protein trans-splicing in vivo and in vitro. *Nat. Protoc.* **5**, 574-587 (2010).

94 Hiraga, K. et al., Selection and structure of hyperactive inteins: peripheral changes relayed to the catalytic center. *J. Mol. Biol.* **393**, 1106-1117 (2009).

95 Appleby-Tagoe, J. H. et al., Highly efficient and more general cis- and trans-splicing inteins through sequential directed evolution. *J. Biol. Chem.* **286**, 34440-34447 (2011).

96 Dassa, B., Amitai, G., Caspi, J., Schueler-Furman, O., and Pietrokovski, S., Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations. *Biochemistry* **46**, 322-330 (2007).

97 Chen, L., Zhang, Y., Li, G., Huang, H., and Zhou, N., Functional characterization of a naturally occurring trans-splicing intein from Synechococcus elongatus in a mammalian cell system. *Anal. Biochem.* **407**, 180-187 (2010).

98 Shah, N. H., Vila-Perelló, M., and Muir, T. W., Kinetic control of one-pot trans-splicing reactions by using a wild-type and designed split intein. *Angew. Chem. Int. Ed.* **50**, 6511-6515 (2011).

99 Du, Z. et al., Backbone dynamics and global effects of an activating mutation in minimized Mtu RecA inteins. *J. Mol. Biol.* **400**, 755-767 (2010).

100 Busche, A. E. L. et al., Segmental isotopic labeling of a central domain in a multidomain protein by protein trans-splicing using only one robust DnaE intein. *Angew. Chem. Int. Ed.* **48**, 6128-6131 (2009).

101 Dhar, T. and Mootz, H. D., Modification of transmembrane and GPI-anchored proteins on living cells by efficient protein trans-splicing using the Npu DnaE intein. *Chem. Commun.* **47**, 3063-3065 (2011).

102 Mills, K. V. and Perler, F. B., The mechanism of intein-mediated protein splicing: variations on a theme. *Protein Pept. Lett.* **12**, 751-755 (2005).

103 Olsen, S. K., Capili, A. D., Lu, X., Tan, D. S., and Lima, C. D., Active site remodelling accompanies thioester bond formation in the SUMO E1. *Nature* **463**, 906-912 (2010).

104 Wu, Y.-W. et al., Membrane targeting mechanism of Rab GTPases elucidated by semisynthetic protein probes. *Nat. Chem. Biol.* **6**, 534-540 (2010).

105 Lu, W. et al., Split intein facilitated tag affinity purification for recombinant proteins with controllable tag removal by inducible auto-cleavage. *J . Chromatogr. A* **1218**, 2553-2560 (2011).

106 Volkmann, G., Sun, W., and Liu, X.-Q., Controllable protein cleavages through intein fragment complementation. *Protein Sci.* **18**, 2393-2402 (2009).

107 Xu, M. Q., Paulus, H., and Chong, S., Fusions to self-splicing inteins for protein purification. *Meth. Enzymol.* **326**, 376-418 (2000).

108 Villain, M., Gaertner, H., and Botti, P., Native Chemical Ligation with aspartic and glutamic acids as C-terminal residues: Scope and limitations. *Eur. J. Org. Chem.* **2003**, 3267-3272 (2003).

109 McGinty, R. K., Chatterjee, C., and Muir, T. W., Semisynthesis of ubiquitylated proteins. *Meth. Enzymol.* **462**, 225-243 (2009).

110 Wang, Z. et al., Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40**, 897-903 (2008).

111 Alley, S. C., Okeley, N. M., and Senter, P. D., Antibody-drug conjugates: targeted drug delivery for cancer. *Curr. Opin. Chem. Biol.* **14**, 529-537 (2010).

112 Webb, S., Pharma interest surges in antibody drug conjugates. *Nat. Biotechnol.* **29**, 297-298 (2011).

113 Mohlmann, S., Bringmann, P., Greven, S., and Harrenga, A., Site-specific modification of ED-B-targeting antibody using intein-fusion technology. *BMC Biotechnol.* **11**, 76 (2011).

114 Scheck, R. A. and Francis, M. B., Regioselective labeling of antibodies through N-terminal transamination. *ACS Chem. Biol.* **2**, 247-251 (2007).

115 Hudak, J. E. et al., Synthesis of heterobifunctional protein fusions using copper-free click chemistry and the aldehyde tag. *Angew. Chem. Int. Ed.* **51**, 4161-4165 (2012).

116 Casi, G., Huguenin-Dezot, N., Zuberbühler, K., Scheuermann, J., and Neri, D., Site-specific traceless coupling of potent cytotoxic drugs to recombinant antibodies for pharmacodelivery. *J. Am. Chem. Soc.* **134**, 5887-5892 (2012).

117 Kazane, S. A. et al., Site-specific DNA-antibody conjugates for specific and sensitive immuno-PCR. *Proc. Natl. Acad. Sci. USA* **109**, 3731-3736 (2012).

118 Witte, M. D. et al., Preparation of unnatural N-to-N and C-to-C protein fusions. *Proc. Natl. Acad. Sci. USA* **109**, 11993-11998 (2012).

119 Barbuto, S. et al., Induction of innate and adaptive immunity by delivery of poly dA:dT to dendritic cells. *Nat. Chem. Biol.* **9**, 250-256 (2013).

120 Bonifaz, L. C. et al., In vivo targeting of antigens to maturing dendritic cells via the DEC-205 receptor improves T cell vaccination. *J. Exp. Med.* **199**, 815-824 (2004).

121 Idoyaga, J. et al., Cutting edge: langerin/CD207 receptor on dendritic cells mediates efficient antigen presentation on MHC I and II products in vivo. *J. immunol.* **180**, 3647-3650 (2008).

122 Shemella, P. T. et al., Electronic structure of neighboring extein residue modulates intein C-terminal cleavage activity. *Biophys. J.* **100**, 2217-2225 (2011).

123 Shah, N. H., Dann, G. P., Vila-Perelló, M., Liu, Z., and Muir, T. W., Ultrafast protein splicing is common among cyanobacterial split inteins: Implications for protein engineering. *J. Am. Chem. Soc.* **134**, 11338-11341 (2012).

124 Vila-Perelló, M. et al., Streamlined expressed protein ligation using split inteins. *J. Am. Chem. Soc.* **135**, 286-292 (2013).

125 Jagadish, K. et al., Expression of fluorescent cyclotides using protein trans-splicing for easy monitoring of cyclotide-protein interactions. *Angew. Chem. Int. Ed.* **52**, 3126-3131 (2013).

126 Zhang, Y. et al., Development of a novel DnaE intein-based assay for quantitative analysis of G-protein-coupled receptor internalization. *Anal. Biochem.* **417**, 65-72 (2011).

127 Wong, S., Mills, E., and Truong, K., Simultaneous assembly of two target proteins using split inteins for live cell imaging. *Protein Eng. Des. Sel.* (2012).

128 Cheriyan, M., Pedamallu, C. S., Tori, K., and Perler, F., Faster protein splicing with the Nostoc punctiforme DnaE intein using non-native extein residues. *J. Biol. Chem.* (2013).

129 Schreiber, G. and Fersht, A. R., Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* **3**, 427-431 (1996).

130 Selzer, T., Albeck, S., and Schreiber, G., Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.* **7**, 537-541 (2000).

131 Graddis, T. J., Myszka, D. G., and Chaiken, I. M., Controlled formation of model homo- and heterodimer coiled coil polypeptides. *Biochemistry* **32**, 12664-12671 (1993).

132 O'Shea, E. K., Lumb, K. J., and Kim, P. S., Peptide 'Velcro': design of a heterodimeric coiled coil. *Curr. Biol.* **3**, 658-667 (1993).

133 Gunasekaran, K. et al., Enhancing antibody Fc heterodimer formation through electrostatic steering effects: applications to bispecific molecules and monovalent IgG. *J. Biol. Chem.* **285**, 19637-19646 (2010).

134 Zhang, Z. et al., Programmable one-pot oligosaccharide synthesis. *J. Am. Chem. Soc.* **121**, 734-753 (1999).

135 Bang, D., Pentelute, B. L., and Kent, S. B. H., Kinetically controlled ligation for the convergent chemical synthesis of proteins. *Angew. Chem. Int. Ed.* **45**, 3985-3988 (2006).

136 Hassa, P. O., Haenni, S. S., Elser, M., and Hottiger, M. O., Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going? *Microbiol. Mol. Biol. Rev.* **70**, 789-829 (2006).

137 Martin, S. A., Lord, C. J., and Ashworth, A., DNA repair deficiency as a therapeutic target in cancer. *Curr. Opin. Genet. Dev.* **18**, 80-86 (2008).

138 Tao, Z., Gao, P., and Liu, H.-w., Identification of the ADP-ribosylation sites in the PARP-1 automodification domain: analysis and implications. *J. Am. Chem. Soc.* **131**, 14258-14260 (2009).

139 Knight, M. I. and Chambers, P. J., Production, extraction, and purification of human poly(ADP-ribose) polymerase-1 (PARP-1) with high specific activity. *Protein Expr. Purif.* **23**, 453-458 (2001).

140 Gagnon, S. N. and Desnoyers, S., Single amino acid substitution enhances bacterial expression of PARP-1(D214A). *Mol. Cell. Biochem.* **243**, 15-22 (2003).

141 Tao, Z., Gao, P., Hoffman, D. W., and Liu, H.-W., Domain C of human poly(ADP-ribose) polymerase-1 is important for enzyme activity and contains a novel zinc-ribbon motif. *Biochemistry* **47**, 5804-5813 (2008).

142 Lilyestrom, W., van der Woerd, M. J., Clark, N., and Luger, K., Structural and biophysical studies of human PARP-1 in complex with damaged DNA. *J. Mol. Biol.* **395**, 983-994 (2010).

143 Simonin, F. et al., The carboxyl-terminal domain of human poly(ADP-ribose) polymerase. Overproduction in Escherichia coli, large scale purification, and characterization. *J. Biol. Chem.* **268**, 13454-13461 (1993).

144 Pinnola, A., Naumova, N., Shah, M., and Tulin, A. V., Nucleosomal core histones mediate dynamic regulation of poly(ADP-ribose) polymerase 1 protein binding to chromatin and induction of its enzymatic activity. *J. Biol. Chem.* **282**, 32511-32519 (2007).

145 Southan, G. and Szabo, C., Poly(ADP-ribose) polymerase inhibitors. *Curr. Med. Chem.* **10**, 321-340 (2003).

146 Messner, S. et al., PARP1 ADP-ribosylates lysine residues of the core histone tails. *Nucleic Acids Res.* **38**, 6350-6362 (2010).

147 Southworth, M. W. et al., Control of protein splicing by intein fragment reassembly. *EMBO J.* **17**, 918-926 (1998).

148 Zheng, Y., Wu, Q., Wang, C., Xu, M.-Q., and Liu, Y., Mutual synergistic protein folding in split intein. *Biosci. Rep.* **32**, 433-442 (2012).

149 Sorci, M. et al., Oriented covalent immobilization of antibodies for measurement of intermolecular binding forces between zipper-like contact surfaces of split inteins. *Anal. Chem.* **85**, 6080-6088 (2013).

150 Shah, N. H. and Muir, T. W., Split inteins: Nature's protein ligases. *Isr. J. Chem.* **51**, 854-861 (2011).

151 Shah, N. H., Eryilmaz, E., Cowburn, D., and Muir, T. W., Extein residues play an intimate role in the rate-limiting step of protein trans-splicing. *J. Am. Chem. Soc.* **135**, 5839-5847 (2013).

152 Farrow, N. A. et al., Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. *Biochemistry* **33**, 5984-6003 (1994).

153 Muller-Späth, S. et al., Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. USA* **107**, 14609-14614 (2010).

154 Semisotnov, G. V. et al., Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe. *Biopolymers* **31**, 119-128 (1991).

155 Pettersen, E. F. et al., UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605-1612 (2004).

156 Millet, O. et al., The static magnetic field dependence of chemical exchange linebroadening defines the NMR chemical shift time scale. *J. Am. Chem. Soc.* **122**, 2867-2887 (2000).

157 Wishart, D. S., Sykes, B. D., and Richards, F. M., The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* **31**, 1647-1651 (1992).

158 Johnson, B. A., Using NMRView to visualize and analyze the NMR spectra of macromolecules. *Methods Mol. Biol.* **278**, 313-352 (2004).

159 Shoemaker, B. A., Portman, J. J., and Wolynes, P. G., Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. USA* **97**, 8868-8873 (2000).

160 Carvajal-Vallejos, P., Pallissé, R., Mootz, H. D., and Schmidt, S. R., Unprecedented rates and efficiencies revealed for new natural split inteins from metagenomic sources. *J. Biol. Chem.* **287**, 28686-28696 (2012).

161 Sun, W., Yang, J., and Liu, X.-Q., Synthetic two-piece and three-piece split inteins for protein trans-splicing. *J. Biol. Chem.* **279**, 35281-35286 (2004).

162 Whitton, B. A. and Potts, M. eds., *The ecology of cyanobacteria: Their diversity in time an space*. (Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000).

163 Vioque, A., Transformation of cyanobacteria. *Adv. Exp. Med. Biol.* **616**, 12-22 (2007).

164 Castenholz, R. W., Culturing methods for cyanobacteria. *Meth. Enzymol.* **167**, 68-93 (1988).

165 Bryksin, A. V. and Matsumura, I., Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. *BioTechniques* **48**, 463-465 (2010).

166 Gasteiger, E. et al., Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*, 571-607 (2005).

167 Schwarzer, D., Ludwig, C., Thiel, I. V., and Mootz, H. D., Probing intein-catalyzed thioester formation by unnatural amino acid substitutions in the active site. *Biochemistry* **51**, 233-242 (2012).

168 Kun, E., Kirsten, E., Mendeleyev, J., and Ordahl, C. P., Regulation of the enzymatic catalysis of poly(ADP-ribose) polymerase by dsDNA, polyamines, Mg2+, Ca2+, histones H1 and H3, and ATP. *Biochemistry* **43**, 210-216 (2004).

169 Moyle, P. M. and Muir, T. W., Method for the synthesis of mono-ADP-ribose conjugated peptides. *J. Am. Chem. Soc.* **132**, 15878-15880 (2010).

170 Delaglio, F. et al., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277-293 (1995).

171 Sattler, M., Schleucher, J., and Griesinger, C., Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Mag. Res. Sp.* **34**, 93-158 (1999).

172 Case, D. A. et al., AMBER 12. *AMBER 12*. **University of California, San Francisco** (2012).

173 Lindorff-Larsen, K. et al., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950-1958 (2010).

174 Jorgensen, W. L. and Madura, J. D., Quantum and statistical mechanical studies of liquids. 25. Solvation and conformation of methanol in water. *J. Am. Chem. Soc* **105**, 1407-1413 (1983).

175 Grest, G. and Kremer, K., Molecular dynamics simulation for polymers in the presence of a heat bath. *Phys. Rev. A* **33**, 3628-3631 (1986).

176 Darden, T., York, D., and Pedersen, L., Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089-10092 (1993).

177 Lippert, R. A. et al., A common, avoidable source of error in molecular dynamics integrators. *J. Chem. Phys.* **126**, 046101 (2007).

178 Gough, J., Karplus, K., Hughey, R., and Chothia, C., Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903-919 (2001).

179 Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D., Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531 (1994).

180 Eddy, S. R., Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).

181 Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E., WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188-1190 (2004).

182 Kyte, J. and Doolittle, R. F., A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105-132 (1982).