Published in final edited form as:

J Am Chem Soc. 2012 July 18; 134(28): 11338–11341. doi:10.1021/ja303226x.

Ultrafast Protein Splicing is Common Among Cyanobacterial Split Inteins: Implications for Protein Engineering

Neel H. Shah, Geoffrey P. Dann, Miquel Vila-Perelló, Zhihua Liu, and Tom W. Muir*

Department of Chemistry, Princeton University, 325 Frick Chemistry Laboratory, Princeton, New Jersey 08544, USA

Abstract

We describe the first systematic study of a family of inteins, the split DnaE inteins from cyanobacteria. By measuring *in vivo* splicing efficiencies and *in vitro* kinetics, we demonstrate that several inteins can catalyze protein *trans*-splicing in tens of seconds, rather than hours, as is commonly observed for this autoprocessing protein family. Furthermore, we show that when artificially fused, these inteins can be used to rapidly generate protein α -thioesters for expressed protein ligation. This comprehensive survey of split inteins provides indispensable information for the development and improvement of intein-based tools for chemical biology.

Protein splicing is a post-translational process catalyzed by a family of proteins known as inteins.¹ During this process, an intein domain catalyzes its own excision from a larger precursor protein and simultaneously ligates the two flanking polypeptide sequences (exteins) together. While most inteins catalyze splicing *in cis*, a small subset of these proteins exist as naturally fragmented domains that are separately expressed but rapidly associate and catalyze splicing *in trans* (Figures 1a, S1). Given their capacity to make and break peptide bonds (inteins can be considered protein ligases), both *cis*- and *trans*-splicing inteins have found widespread use as chemical biological tools.²

Despite the growing use of inteins in chemical biology, their practical utility has been constrained by two common characteristics of the family, namely (i) slow kinetics and (ii) context dependent efficiency with respect to the immediate flanking extein sequences.^{3,4} 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.⁵ Furthermore, this intein was slightly more tolerant of sequence variation at the critical +2 C-extein residue than other characterized inteins (Figure 1a).⁶

We became interested in the apparently unique properties of Npu and sought to determine whether other homologous split inteins also catalyze rapid *trans*-splicing, perhaps with greater C-extein tolerance. Of the roughly 600 inteins currently catalogued,⁷ less than 5% are split in-teins, mostly from a family known as the cyanobacterial split DnaE inteins⁸ (Figure S2, Table S1). Surprisingly, only six of these, including Npu, have been experimentally analyzed to any extent,^{6,9,10} and only Npu and its widely-studied, low-efficiency ortholog from *Synechocystis* species PCC6803 (Ssp) have been rigorously characterized *in vitro*.^{5,11}

Corresponding Author: muir@princeton.edu.

Supporting Information: Full methods and experimental data. This information is available free of charge via the Internet at <http://pubs.acs.org>.

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.^{12,13} In this assay, the two fragments of a split intein are co-expressed in *E. coli* as fusions to a fragmented aminoglycoside phosphotransferase (Kan^R) enzyme. Upon *trans*-splicing, the active enzyme is assembled, and the bacteria become resistant to the antibiotic kanamycin (Figures 1a, S3a). More active inteins confer greater kana-mycin resistance and thus have a higher IC₅₀ value for bacterial growth as a function of kanamycin concentration. Importantly, this assay can be carried out in the background of varying local C-extein sequences without significantly perturbing the dynamic range. Since all DnaE inteins splice the same local extein sequences in their endogenous context, we initially carried out our screen in a wild-type C-extein background (CFN) within the Kan^R enzyme. As expected, bacteria expressing the Npu intein had a high relative IC₅₀, whereas clones expressing Ssp showed poor resistance to kanamycin. Remarkably, more than half of the DnaE inteins showed splicing efficiency comparable to Npu *in vivo* at 30°C (Figures 1b, S4-S6).

To confirm that the high IC₅₀ 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 (Figures S7-S12 and Tables S2, S3). Importantly, we preserved the endogenous local extein residues as linkers between the extein domains and intein fragments to recapitulate a wild-typelike splicing context (Figure S3b). Cognate intein fragments were mixed at 1μM, and the formation of the Ub-SUMO spliced product at 30°C and 37°C was monitored by gel electrophoresis. These assays validated that the new in-teins with high-activity *in vivo* could catalyze *trans*-splicing *in vitro* in tens of seconds, substantially faster than Ssp (Figures 2a, S13-S15 and Tables S4, S5). 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 (Figure 2b), again in contrast to Ssp (Figure 2c).

Next, we investigated the tolerance of the split inteins to C-extein sequence variation. Previously, we and others have noted the sensitivity of DnaE inteins to changes at the +2 position in the C-extein.^{6,12} Thus, we analyzed all the split DnaE inteins in the presence of a +2 glycine (CGN), glutamic acid (CEN), or arginine (CRN) in our *in vivo* screening assay (Figures 1b, S4-S6). Like Npu and Ssp, most of the inteins showed a dramatic decrease in activity in the presence of all three +2 mutations. 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. To more accurately assess the magnitude of the effect of C-extein mutations on *trans*-splicing, we analyzed the Npu, Cra(CS505), and Cwa inteins *in vitro* in the presence of a +2 glycine (Figures S16-S20). All three of these reactions were characterized by rapid accumulation of thioester intermediates, which slowly resolved over tens of minutes into the spliced product and the N-extein cleavage product. 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.¹² It is noteworthy that the Cra(CS505) and Cwa inteins showed greater C-extein promiscuity *in vivo*, while Ssp(PCC7002) did not tolerate any of the mutations we tested. This demonstrates that subtle sequence variation between split inteins can afford differential promiscuity. Thus, this property may be further optimized through directed evolution¹² or rational design.

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. Interestingly, the key catalytic residues involved in splicing are conserved across the entire family (Figure S2). 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 S21). 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, was inactive with the wild-type “CFN” C-extein motif *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 3a). Furthermore, this position is close to the intein active site, where an extra nucleophile may facilitate undesirable side reactions (Figure 3b). Gratifyingly, mutating this cysteine to glycine reinstated high activity in the Aha intein whilst the reverse mutation destroyed the splicing activity of Npu (Figures 3c, S23a), validating the predictive capacity of our data.

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 (Figures 3a, S22). 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. Consistent with our hypothesis, several point mutations increased the activity of Ssp *in vivo* (Figures 3e, S23b). While the specific roles of these residues are not explicitly clear, especially given that they lie outside of the active site (Figure 3d), their locations on the intein fold¹⁴ may provide some insights into their function (Figure S24). 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. The glutamate at position 89 is involved in an intimate ion cluster that we have previously shown is important for stabilizing the split intein complex.¹³ 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 activating point mutations in other inteins.^{15,16}

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.^{13,17,18} 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 (Figures S7, S8). 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 S25).

The most widely used intein-based technology, expressed protein ligation, exploits *cis*-acting inteins to generate recombinant protein α -thioester derivatives.² In principle, any split intein can be artificially fused and then utilized as a *cis*-splicing intein in this application (1 in Figure 4a). 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. 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 (Figures 4b and S27). By contrast, MESNa thiolysis of the commonly used MxeGyrA intein was not complete even after one day under identical conditions. 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 semisynthetic protein **6** (Figure 4c). 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 S29). These results demonstrate that fused versions of split DnaE inteins will be of general utility for protein semisynthesis.

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 4a). This thioester intermediate is generally thought to be transiently populated in protein splicing, and to our knowledge, it has never been directly observed.¹ Surprisingly, when analyzing the ubiquitin-DnaE intein fusions by reverse phase HPLC, we often observed two major peaks and a third minor peak, all bearing the same mass (Figure S30). The relative abundance of these species could be modulated by unfolding the proteins or by changes in pH, and the two major species were almost equally populated from pH 4-6 (Figure 4d). The major peaks most likely correspond to the precursor amide, **1**, and the linear thioester, **2**, and we speculate that the minor peak is the tetrahedral oxythiazolidine intermediate. Importantly, only a single HPLC peak was seen for the ubiquitin-MxeGyrA fusion under identical conditions (Figure S30). These observations, along with the enhanced thiolysis rates, strongly support the notion that these DnaE inteins have a hyper-activated N-terminal splice junction.

In this study, we systematically characterized splicing activities in an entire family of split inteins. We demonstrated that ultrafast protein *trans*-splicing is the norm, rather than the exception, in this family. Furthermore, we showed that different split inteins have varying degrees of tolerance for C-extein mutations, suggesting that traceless protein splicing may be attainable by modestly engineering any highly active intein. We also illustrated that a thorough comparison of the activities of a small family of homologous proteins can be used to identify important non-catalytic positions that modulate activity. Finally, by artificially fusing split DnaE intein fragments, we generated new constructs for the efficient synthesis of protein α -thioesters used in expressed protein ligation. These results will guide the development of improved protein chemistry technologies and should lay the groundwork towards a more fundamental understanding of efficient protein splicing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank members of the Muir laboratory for valuable discussions. This work was supported by grants from the US National Institutes of Health (GM086868).

References

1. Mills KV, Perler FB. Protein Pept Lett. 2005; 12:751–5. [PubMed: 16305544]
2. Vila-Perelló M, Muir TW. Cell. 2010; 143:191–200. [PubMed: 20946979]

3. Southworth M, Amaya K, Evans T, Xu M, Perler F. *Biotechniques*. 1999; 27:110–120. [PubMed: 10407673]
4. Amitai G, Callahan BP, Stanger MJ, Belfort G, Belfort M. *Proc Natl Acad Sci USA*. 2009; 106:11005–10. [PubMed: 19541659]
5. Zettler J, Schütz V, Mootz HD. *FEBS Lett*. 2009; 583:909–14. [PubMed: 19302791]
6. Iwai H, Züger S, Jin J, Tam PH. *FEBS Lett*. 2006; 580:1853–8. [PubMed: 16516207]
7. Perler FB. *Nucleic Acids Res*. 2002; 30:383–4. [PubMed: 11752343]
8. Caspi J, Amitai G, Belenkiy O, Pietrokovski S. *Mol Microbiol*. 2003; 50:1569–77. [PubMed: 14651639]
9. Dassa B, Amitai G, Caspi J, Schueler-Furman O, Pietrokovski S. *Biochemistry*. 2007; 46:322–330. [PubMed: 17198403]
10. Chen L, Zhang Y, Li G, Huang H, Zhou N. *Anal Biochem*. 2010; 407:180–7. [PubMed: 20727340]
11. Martin DD, Xu MQ, Evans TC. *Biochemistry*. 2001; 40:1393–402. [PubMed: 11170467]
12. Lockless SW, Muir TW. *Proc Natl Acad Sci USA*. 2009; 106:10999–1004. [PubMed: 19541616]
13. Shah NH, Vila-Perelló M, Muir TW. *Angew Chem Int Ed Engl*. 2011; 50:6511–5. [PubMed: 21656885]
14. Oeemig JS, Aranko AS, Djupsjöbacka J, Heinämäki K, Iwai H. *FEBS Lett*. 2009; 583:1451–1456. [PubMed: 19344715]
15. Du Z, Liu Y, Ban D, Lopez MM, Belfort M, Wang C. *J Mol Biol*. 2010; 400:755–67. [PubMed: 20562025]
16. Appleby-Tagoe JH, Thiel IV, Wang Y, Wang Y, Mootz HD, Liu XQ. *J Biol Chem*. 2011; 286:34440–7. [PubMed: 21832069]
17. Busche AEL, Aranko AS, Talebzadeh-Farooji M, Bernhard F, Dötsch V, Iwai H. *Angew Chem Int Ed Engl*. 2009; 48:6128–31. [PubMed: 19591176]
18. Dhar T, Mootz HD. *Chem Commun*. 2011; 47:3063–5.

Figure 1.

Trans-splicing of split DnaE inteins. (a) Scheme depicting protein *trans*-splicing of the KanR protein with a variable local C-extein sequence. (b) *In vivo* relative *trans*-splicing efficiencies at 30°C with the endogenous “CFN” C-extein sequence and exogenous “CGN”, “CEN”, and “CRN” sequences. IC₅₀ values ($\pm$ SE, $n = 3-4$) are normalized to the intact KanR proteins with the corresponding tri-peptide.

Figure 2.

In vitro trans-splicing reactions. Indicated split intein pairs fused to model exteins Ub and SUMO (Ub-Int^N and Int^C-SUMO) were mixed at 30 °C or 37 °C, and the formation of products was monitored over time by gel electrophoresis. (a) Half-lives were extracted from the reaction progress curves fit to a first-order rate equation ($\pm$ SE, $n = 3$). Coomassie-stained SDS-PAGE gels showing (b) fast Ava splicing at 37°C and (c) inefficient Ssp splicing at 37°C.

Figure 3. 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 highlighting the proximity of position 120 to the terminal catalytic residues C1 and N137. (c) *In vivo* analysis of the C120G mutation in the Aha intein ($\pm$ SD, $n = 3$). (d) Rendering of the Npu structure highlighting key catalytic residues (orange sticks) and important non-catalytic positions (green spheres) that modulate Ssp activity. (e) *In vivo* analysis of Ssp-to-Npu point mutations that improve Ssp activity ($\pm$ SD, $n = 4$). Note, all residue numberings correspond to the relevant positions on Npu as defined by the NMR structure (PDB: 2KEQ).

Figure 4.

Engineered versions of DnaE inteins support efficient expressed protein ligation. (a) Scheme showing the formation of the linear thioester intermediate and its use to generate a protein α -thioester for EPL. (b) Coomassie-stained SDS-PAGE gel depicting MESNa thiolysis of ubiquitin from a fused AvaDnaE intein to yield the Ub-MES thioester, **4**. (c) Fluorescent SDS-PAGE gels showing the formation of the Ub-CGK(Fluorescein) ligated product (**6**) from one-pot thiolysis and native chemical ligation reactions using the inteins indicated. (d) RP-HPLC chromatographs showing pH dependence of precursor amide (**1**) and linear thioester (**2**). A third minor peak is indicated by *.