Published in final edited form as:

J Am Chem Soc. 2018 September 19; 140(37): 11791–11799. doi:10.1021/jacs.8b07334.

An Atypical Mechanism of Split Intein Molecular Recognition and Folding

Adam J. Stevens^{1,#}, Giridhar Sekar^{2,#}, Josef A. Gramespacher¹, David Cowburn^{2,*}, Tom W. Muir^{1,*}

¹Department of Chemistry, Princeton University, Frick Laboratory, Princeton, New Jersey 08544, United States

²Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461, United States

Abstract

Split inteins associate to trigger protein splicing *in trans*, a posttranslational modification in which protein sequences fused to the intein pair are ligated together in a traceless manner. Recently, a family of naturally split inteins has been identified that is split at a non-canonical location in the primary sequence. These atypically split inteins show considerable promise in protein engineering applications, however the mechanism by which they associate is unclear and must be different from that of previously characterized canonically split inteins due to unique topological restrictions. Here we use a consensus design strategy to generate an atypical split intein pair (Cat) that has greatly improved activity and is amenable to detailed biochemical and biophysical analysis. Guided by the solution structure of Cat, we show that the association of the fragments involves a disorder-to-order structural transition driven by hydrophobic interactions. This molecular recognition mechanism satisfies the topological constraints of the intein fold and, importantly, ensures that premature chemistry does not occur prior to fragment complementation. Our data lead a common blueprint for split intein complementation in which localized structural rearrangements are used to drive folding and regulate protein-splicing activity.

GRAPHICAL ABSTRACT

*To whom correspondence should be addressed: cowburn@cowburnlab.org; muir@princeton.edu.

#These authors contributed equally to this work

Introduction

An intein is an intervening protein domain that undergoes a posttranslational autoprocessing event called protein splicing in which it excises itself from a host protein while tracelessly ligating its flanking polypeptide sequences (exteins) to form a native peptide bond.^{1,2} Most inteins are found as contiguous domains embedded within a single gene and splice *in cis*. However, some exist naturally in split form, whereby each intein fragment is encoded on a separately expressed gene and must first associate prior to splicing *in trans* (Figure 1a).^{3–5} These split inteins are commonly applied as tools in protein engineering, and are especially amenable to use in the cellular environment due to their highly specific recognition and unique activity.⁶

Split intein association is characterized by the coupling of molecular recognition to activation of splicing.^{7–10} Prior to assembly, autoprocessing activity remains negligible, thereby mitigating the deleterious effects of side reactions such as N- and C-terminal cleavage of extein sequences from the intein fragments. However upon association, protein *trans*-splicing (PTS) proceeds efficiently to ensure production of the spliced product. Previous efforts have sought to determine the mechanism of this precisely tuned association process both to elucidate native split intein function and inform the design of intein-based technologies.^{7–9,11} For the split inteins found in the α subunit of DNA Polymerase III (DnaE) from the cyanobacteria *Synechocystis sp.* strain PCC6803 (Ssp) and *Nostoc Punctiforme* (Npu), fragment assembly was shown to exploit a disorder to order structural transition to produce the topologically intertwined Hedgehog/Intein (HINT) domain.^{7,8,12} Alternatively, a split intein from the DNA polymerase of *Nanoarchaeum equitans* (Neq) employs an extensive structural rearrangement of a pre-folded Neq N-intein upon C-intein interaction.¹³ Each of these two mechanisms provides precise control of assembly through coordinated structural rearrangements.

Neq, Npu, and Ssp are all fragmented at the canonical intein split site, which also serves as the standard position of homing endonuclease insertion in contiguous inteins.¹ This site is proximal to the C-terminus, generating N- and C-inteins of approximately 100 and 35 amino acids respectively (Figure 1b). However, split inteins with atypical split sites have recently

been reported from metagenomic sequencing data and characterized *in vitro* (Figure 1b).^{14,15} These atypically split inteins contain N-inteins of approximately 30 amino acids, making them accessible to solid phase peptide synthesis and suitable tools for the N-terminal modification of recombinant proteins.^{16,17} Like canonical split inteins, atypically split inteins must be capable of molecular recognition through a specific and efficient association event that results in the formation of a folded HINT domain. Although the HINT domain displays pseudo two-fold symmetry, the respective C- and N- fragments of canonically and atypically split inteins do not mirror one another in a structural sense (Figure 1c). As a consequence, the topological requirements for association and assembly of atypical split inteins are unique: rather than forming an intertwined structure, the atypical split site necessitates threading of the N-extein through the C-intein during assembly (Figure 1c). Canonical and atypical split inteins thus serve as an intriguing example of proteins within a common family evolved to carry out the same function (association and protein splicing), but facing distinct topological obstacles. In this study, we apply the design of a consensus intein possessing superior thermodynamic stability to determine the molecular basis of atypical split intein assembly and activity.

Results

Design of a consensus atypical split intein with enhanced stability and activity

In order to determine the mechanism of fragment association, we first attempted to isolate an atypically split intein with minimal extein residues. Both naturally occurring atypical split inteins whose splicing rates have been characterized *in vitro* were identified within the T4-bacteriophage-type DNA-packaging terminase large subunit (TerL) from metagenomic sequencing data.^{14,15} The first, from the saline meromictic Ace Lake in Antarctica (AceL), exhibits an optimal splicing rate at 8 °C ($t_{1/2}$ = 7 min). In addition, directed evolution found stabilizing mutations within AceL (AceL*) that increase activity at 37 °C ($t_{1/2}$ = 6 min).¹⁵ The second characterized atypical split intein was sequenced in a sample collected from Punta Cormorant in the global ocean sampling project (GOS) and splices at an optimal temperature of 30 °C ($t_{1/2}$ = 3 min).¹⁴ Unfortunately, we were unable to purify soluble GOS^N (i.e. the N-terminal GOS intein fragment), GOS^C, or AceL*^C from expression in *E. coli* without the use of large stabilizing extein proteins (Figure S1). Subsequent efforts to extract atypically split inteins lacking solubilizing exteins from the insoluble inclusion body fraction with chaotropic agents were unsuccessful in our hands due to aggregation issues while refolding.

Consensus design is a protein engineering strategy that utilizes evolutionary information from homologous protein sequences to predict stabilizing mutations and has previously been applied to generate a highly active and thermostable naturally split DnaE intein (Cfa).^{18–21} Seeking to engineer an atypically split intein amenable to *in vitro* structural characterization, we designed a consensus atypical (Cat) TerL intein from multiple sequence alignments (MSA) of TerL^N and TerL^C inteins discovered from BLAST searches of metagenomic sequencing information in the JGI and NCBI databases (Table S1).^{22–24} Both Cat^N (60%) and Cat^C (64%) contain high sequence similarity to AceL*^N and AceL*^C respectively, with the nonidentical residues spread throughout the primary sequence (Figure 2a). The Cat intein

pair was isolated fused to model exteins to measure its *in vitro trans*-splicing activity (Table S2). Cat exhibits ultrafast splicing activity ($t_{1/2} = 59$ s at 30 °C) and consistently outperforms AceL* across an array of temperatures (Figure 2b, 2c). Moreover, Cat remains active at 50 °C, a temperature at which AceL* fails to splice. PTS was also measured in the presence of chaotropic agents, which are often utilized to solubilize aggregation-prone extein fragments.¹ Cat displays enhanced chaotropic stability and can splice in both 2 M and 4 M urea (Figure 2d, Table S3), while AceL* is inactive under both of these conditions. The accelerated splicing rates and activity under adverse conditions establish Cat as the fastest and most robust atypical split intein reported to date, and it should therefore serve as a tool for the synthetic N-terminal modification of proteins.

Fragment assembly drives a disorder to order structural transition

To investigate the association process of atypical split inteins, Cat^N and Cat^C bearing minimal exteins were expressed in isotopically enriched media (¹⁵N, ¹³C), purified, and analyzed by nuclear magnetic resonance (NMR) spectroscopy. Note, these constructs also included inactivating C1A and N134A mutations to prevent splicing during structural analysis of the complex. The ¹H-¹⁵N HSQC spectrum of Cat^N in isolation displays minimal dispersion along the ¹H dimension, a common phenomenon among disordered proteins and previously observed for Ssp^C and Npu^C (Figure 3a).^{7,8} A stark transition occurs upon addition of unlabeled Cat^C, resulting in a well dispersed ¹H-¹⁵N HSQC spectrum, which is consistent with Cat^N folding (Figure 3a). Furthermore, measurements of ¹H-¹⁵N heteronuclear NOEs, spin-spin relaxation rates, and Cα-Cβ chemical shift perturbation in Cat^N provide additional evidence for a disorder to order transition in Cat^N upon binding Cat^C (Figure S2). The ¹H-¹⁵N HSQC of Cat^C in isolation exhibits far fewer crosspeaks than expected from the number of residues in the protein, a feature present in dynamic proteins that are undergoing chemical exchange and previously observed in both Ssp^N and Npu^N (Figure 3b).^{7,8} Addition of unlabeled Cat^N leads to the appearance of new crosspeaks, which indicates a transition to a more ordered complex (Figure 3b). Although the spectral quality of Cat^C in free form precluded our ability to assign the protein, some crosspeaks overlap those observed in the bound form, which suggests that Cat^C in free and bound form share a partial structural identity.

In line with the NMR studies, analysis by circular dichroism spectroscopy indicates that unbound Cat^N is largely unstructured with some propensity to sample secondary structure, and that both Cat^N and Cat^C inteins undergo a structural transition upon association (Figure 3c). Further evidence for folding upon binding was observed by size exclusion chromatography (SEC), as Cat^C elutes at an earlier time than the bound complex despite having a lower molecular weight (Figure 3d). The SEC elution profile is consistent with a compaction of Cat^C upon binding its cognate intein.

Solution structure of an atypical split intein complex

The isotopically enriched Cat^N and Cat^C proteins were assembled into a complex, and its structure was calculated from distance restraints and dihedral angle constraints obtained from NMR spectroscopy (Supplementary information). The twenty lowest energy conformers obtained from the structure calculation are shown (Figure 4a, PDB ID: 6DSL).

The structure ensemble is precise in all regions of the protein (with the exception of a short solubility tag in Cat^C and the exteins) with a mean backbone RMSD of 1.19 Å to the average structure (Table 1). Residue wise backbone RMSD values of < 0.5 Å were obtained across the structured regions of the protein (Figure S3a, S3b). The structure of Cat is predominantly β -sheet, with the last 8 residues present in the C-terminus of Cat^N being the only α -helix (Figure 4b). It has a horseshoe-like shaped structure that is typical for proteins containing the HINT domain.¹⁰ The structure of Cat is similar to that of DnaE inteins, such as Npu (PDB ID: 2KEQ, RMSD 1.45 Å over 92 aligned C α atoms) and Ssp (PDB ID: 1ZDE, RMSD 1.34 Å over 90 aligned C α atoms) with the notable exception that Npu and Ssp have an additional helix, which is absent in Cat.^{12,25,26}

In the Cat active site, a serine residue (Ser₇₅) replaces the threonine located in the canonical TXXH B-block motif (Figure S3c).^{27,28} The carbonyl oxygen of C1A is proximal to the amide proton (2.4 Å) and the hydroxyl proton (3.7 Å) of Ser₇₅ (Figure 4c). The threonine residue in DnaE inteins adopts a similar conformation, suggesting that Ser₇₅ supplants the role of threonine in assisting the cleavage of the N-terminal scissile peptide bond.²⁵ Another notable feature in the structure is the lack of an F-block histidine (Figure S3c), and therefore resolution of the branched intermediate is likely mediated by the penultimate G-block histidine (His₁₃₃).^{29,30}

Mapping disorder localization in Cat

Limited proteolysis by thermolysin digestion was applied to investigate the distribution of local structure in Cat (Figure 5a). In isolation, Cat^N undergoes rapid degradation, while Cat^C displays slightly greater resistance to proteolysis. The intein complex, however, remains intact after 30 minutes. The variation in protease susceptibility observed is consistent with a largely disordered Cat^N, partially disordered Cat^C, and formation of a globular fold upon binding. We next examined cleavage products (t = 30 min) using electrospray ionization mass spectrometry (ESI-MS) to determine the regions protected from proteolysis, which should correspond to localized structural elements (Figure S4, Table S4). For Cat^N, cut sites appeared to be evenly spread throughout the primary sequence. Conversely, a large portion of Cat^C is resistant to proteolysis. Numerous peaks corresponding to intact fragments centered on residues 57 through 112 were observed, which points to this area as a structured region flanked by disordered N- and C-terminal peptides (Figure 5b). Mapping this model onto the structure of Cat indicates that the disordered N- and C-terminal ends of Cat^C directly interact with Cat^N (Figure 5c). Moreover, key catalytic residues for succinimide formation (Asp₁₁₅, His₁₃₃, and Asn₁₃₄) are present within the disordered region of Cat^C.^{29–31}

Assembly is largely driven by hydrophobic interactions

After examining the structural properties of the Cat fragments in split form, we sought to identify the molecular components that drive association. Although the primary sequences of Cat^N and Cat^C exhibit separation of charge, the binding surface of Cat^N-Cat^C is rich in hydrophobic residues (Figure 6a, b). In the complex, the charged residues of both Cat^N and Cat^C are excluded towards the exterior of the protein while hydrophobic residues are clustered within the binding interface (Figure S5a, S5b). To validate that these hydrophobic

interactions drive complex formation, the effect of buffer ionic strength on fragment association was evaluated using a fluorescence anisotropy-based binding assay (Supplementary information). Cat^N containing an N-terminal fluorescein (Fl-Cat^N) was synthesized by solid phase peptide synthesis, and an increase in fluorescence anisotropy was observed upon association with a SUMO-Cat^C fusion protein (Figure 6c). This increased anisotropy is consistent with an expected increase in rotational correlation time for the Cat complex compared to unbound Cat^N, and was used as a measure of Cat complex formation. Like other split inteins, Cat^N and Cat^C exhibit high binding affinity *in vitro*, with K_d values below 500 pM, which was the limit of detection of the assay (Table S5). Importantly, the binding isotherm for Cat complex formation is minimally perturbed by a change in ionic strength of the buffer, consistent with an association process driven by hydrophobic interactions.

Kinetics of binding between Fl-Cat^N and SUMO-Cat^C were next monitored by stopped-flow fluorescence, and the data was found to be best fit to a double exponential model (Figure S5c). Both determined rate constants (k_{obs1} and k_{obs2}) exhibit concentration dependence leading to a calculated k_{on1} of $(2.80 \pm 0.28) \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$ and k_{on2} of $(0.16 \pm 0.019) \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$ under low salt conditions and k_{on1} of $(2.34 \pm 0.30) \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$ and k_{on2} of $(0.18 \pm 0.016) \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$ under high salt conditions (Figure 6d, Table S6). This model suggests that parallel association events may proceed from distinct conformers of the intein, with subsets of conformers being kinetically distinguishable. Moreover, the observation that both k_{obs1} and k_{obs2} are unperturbed by buffer ionic strength across all measured Cat^C concentrations further suggests that association is largely driven by hydrophobic interactions.

The Extein Dependence of Cat

To date, all characterized inteins exhibit splicing rates dependent on their flanking extein residues.¹ Deviation from the native extein sequence often decelerates splicing and consequently may limit applications of PTS.^{5,32–34} The extein dependence of TerL inteins has yet to be thoroughly characterized, and we therefore sought to identify the sequence preferences of Cat by introducing substitutions that vary charge and steric bulk from the native residues (Figure 7a). Substitutions from the native C-extein, which is Cys₊₁, Glu₊₂, Phe₊₃, were introduced at the +2 and +3 positions and assayed *in vitro* (Figure 7b, Table S7). Cat demonstrates remarkable C-extein promiscuity, splicing with half-lives ranging from 1 to 3 minutes. This broad tolerance to C-extein substitutions is superior even to an engineered version of Npu previously designed to possess promiscuous activity.³⁵ Unlike the tolerance to C-extein substitution, Cat exhibits a stark dependence on the identity of the –1 residue: decreased activity results from inserting alanine ($t_{1/2} = 54 \text{ min}$), glycine ($t_{1/2} = 146 \text{ min}$), or proline ($t_{1/2} = 158 \text{ min}$) at this position (Figure 7c, Table S7). The measured *in vitro* extein dependence is likely explained by interactions observed in the solution structure of the Cat complex. Both Glu₊₂ and Phe₊₃ appear to have minimal contact with active site-catalytic residues, agreeing with the experimentally observed C-extein promiscuity (Figure 7d). Interestingly, Glu₊₂ does contact Asn₁₂₃, which is present in place of an F-block histidine. Conversely, Glu_{–1} directly interacts with Ser₇₅ and His₇₈, two key catalytic residues that

contribute to thioester formation (Figure 7e). N-extein substitutions may therefore directly interfere with the capability of Ser₇₅ and His₇₈ to catalyze protein splicing.

Discussion and Conclusions

In this study, we report the manner by which an atypical split intein assembles and folds into the conserved HINT domain. This mechanism satisfies the unique topological requirements for association imposed by the atypical split site. One such precondition, the required threading of the N-extein through the C-intein, is addressed by coupling fragment association with a disorder-to-order structural transition. Moreover, linking binding to folding provides a thermodynamic drive for the intein association process, which likely contributes to the high affinity between the cognate fragments ($K_d < 500$ pM). Unlike previously characterized naturally split DnaE inteins in which the larger fragment contains two distinct lobes (one structured and one disordered), Cat^C displays three alternating structural regions: a disordered N-terminal region, a central core with local order, and a disordered C-terminal region. This specific architecture is optimal for atypical split inteins, as the disordered regions are proximal to the Cat^N peptide fragment in the bound complex and participate in the hydrophobically driven association process.

Along with overcoming topological barriers to association, the reported mechanism of atypical intein association corresponds to the precise control of catalytic activity necessary for protein *trans*-splicing. Importantly, the disordered C-terminal region of Cat^C encompasses the conserved catalytic residues Asp₁₁₅, His₁₃₃, and Asn₁₃₄, which precludes premature cyclization of the C-terminal Asn₁₃₄ prior to assembly of the active site. Interestingly, the complementary residues in DnaE inteins are present in the disordered C-intein.^{7,8} Furthermore, although Ser₇₅ and His₇₈ are likely catalytic residues and located within the structured central domain of Cat^C, they would be responsible for activation of the N-terminal scissile bond and thus require association with Cat^N to be active. The corresponding residues within the DnaE N-inteins are located within a disordered lobe of the protein, thereby avoiding activating the N-terminal scissile bond in that context.⁷ In considering both the steric and biochemical implications of the reported mechanism of association, a common blueprint for naturally split intein assembly emerges. Prior to association, local structural elements are optimized to limit side reactions and drive interaction between the intein fragments. Upon binding, these intein fragments then undergo a structural rearrangement to form the HINT domain and activate protein splicing.

Beyond the mechanistic implications of atypical split intein association, the Cat intein presented herein should be a powerful tool for protein semisynthesis. Cat is the fastest and most robust atypical split intein reported to date, and it should therefore find immediate utility in the N-terminal modification of expressed proteins.^{16,17,36} This functionality should complement other reported methods for protein N-terminal modification, such as expressed protein ligation, transpeptidase-based ligation strategies, artificially split inteins and various protein chemistry methods.^{36–41} Furthermore, under specific contexts, the extein preferences observed will favor the application of Cat in protein engineering: Cat exhibits promiscuous activity in the presence of substitutions within the C-extein but is highly vulnerable to

mutations within the N-extein. This preference is complementary to the commonly applied naturally split DnaE inteins, which are sensitive to C-extein mutations.^{33,34}

This work also highlights the potential of consensus design to facilitate functional studies of poorly behaved proteins. Although we initially aimed to investigate assembly of previously reported atypically split inteins, challenges obtaining soluble quantities of GOSC and AceL*^C necessitated the identification of a homologue with increased solubility. Rather than screen a series of wild-type TerL homologues for expression, consensus design served as an efficient and rapid means to determine a stable candidate for structural study. We find this strategy especially favorable, as the consensus sequence obtained from an MSA is often more stable than any individual member within the alignment.^{19,42} Lastly, we envision that the mechanism of association and structure of the complex reported in this paper should enable future efforts to engineer enhanced functionality into Cat. Previous work has utilized structure-guided design to generate inteins with increased promiscuity, conditional activity, and truncated fragments, and we envision that similar approaches should be feasible with Cat.^{11,35,43,44}

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Dr. Robert Thompson, Dr. Antony Burton, Dr. Richard Harris, and members of the Muir laboratory for valuable discussions. This work was supported by the U.S. National Institutes of Health (NIH grants R37-GM086868, 1S10OD016305, CO6RR015495, and S10OD016432) and a National Science Foundation Graduate Research Fellowship to AJS (Grant No. DGE-1148900).

References

- (1). Shah NH; Muir TW Chem. Sci 2014, 5, 15.
- (2). Novikova O; Topilina N; Belfort M J Biol Chem 2014, 289, 14490. [PubMed: 24695741]
- (3). Aranko AS; Wlodawer A; Iwai H Protein Eng Des Sel 2014, 27, 263. [PubMed: 25096198]
- (4). Wu H; Hu Z; Liu XQ Proc Natl Acad Sci U S A 1998, 95, 9226. [PubMed: 9689062]
- (5). Iwai H; Zuger S; Jin J; Tam PH FEBS Lett. 2006, 580, 1853. [PubMed: 16516207]
- (6). Vila-Perello M; Muir TW Cell 2010, 143, 191. [PubMed: 20946979]
- (7). Shah NH; Eryilmaz E; Cowburn D; Muir TW J Am Chem Soc 2013, 135, 18673. [PubMed: 24236406]
- (8). Zheng Y; Wu Q; Wang C; Xu MQ; Liu Y Biosci Rep 2012, 32, 433. [PubMed: 22681309]
- (9). Gordo V; Aparicio D; Perez-Luque R; Benito A; Vilanova M; Uson I; Fita I; Ribo M Cell Chem Biol 2018.
- (10). Eryilmaz E; Shah NH; Muir TW; Cowburn D J Biol Chem 2014, 289, 14506. [PubMed: 24695731]
- (11). Gramespacher JA; Stevens AJ; Nguyen DP; Chin JW; Muir TW J Am Chem Soc 2017, 139, 8074. [PubMed: 28562027]
- (12). Oeemig JS; Aranko AS; Djupsjobacka J; Heinamaki K; Iwai H FEBS Lett 2009, 583, 1451. [PubMed: 19344715]
- (13). Gordo V; Aparicio D; Perez-Luque R; Benito A; Vilanova M; Uson I; Fita I; Ribo M Cell Chem Biol 2018, 25, 871. [PubMed: 29754955]
- (14). Bachmann AL; Mootz HD J Biol Chem 2015, 290, 28792. [PubMed: 26453311]

- (15). Thiel IV; Volkmann G; Pietrokovski S; Mootz HD *Angew Chem Int Ed Engl* 2014, 53, 1306. [PubMed: 24382817]
- (16). Bachmann AL; Mootz HD *J Pept Sci* 2017, 23, 624. [PubMed: 28332258]
- (17). Matern JC; Bachmann AL; Thiel IV; Volkmann G; Wasmuth A; Binschik J; Mootz HD *Methods Mol Biol* 2015, 1266, 129. [PubMed: 25560072]
- (18). Porebski BT; Buckle AM *Protein Eng Des Sel* 2016, 29, 245. [PubMed: 27274091]
- (19). Lehmann M; Loch C; Middendorf A; Studer D; Lassen SF; Pasamontes L; van Loon APGM; Wyss M *Protein Engineering* 2002, 15, 403. [PubMed: 12034860]
- (20). Steipe B; Schiller B; Pluckthun A; Steinbacher S *J Mol Biol* 1994, 240, 188. [PubMed: 8028003]
- (21). Stevens AJ; Brown ZZ; Shah NH; Sekar G; Cowburn D; Muir TW *J Am Chem Soc* 2016, 138, 2162. [PubMed: 26854538]
- (22). Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ *J Mol Biol* 1990, 215, 403. [PubMed: 2231712]
- (23). Grigoriev IV; Nordberg H; Shabalov I; Aerts A; Cantor M; Goodstein D; Kuo A; Minovitsky S; Nikitin R; Ohm RA; Otilar R; Poliakov A; Ratnere I; Riley R; Smirnova T; Rokhsar D; Dubchak I *Nucleic Acids Res* 2012, 40, D26. [PubMed: 22110030]
- (24). Tatusova T; Ciufu S; Fedorov B; O'Neill K; Tolstoy I *Nucleic Acids Res* 2014, 42, D553. [PubMed: 24316578]
- (25). Sun P; Ye S; Ferrandon S; Evans TC; Xu MQ; Rao Z *J Mol Biol* 2005, 353, 1093. [PubMed: 16219320]
- (26). Aranko AS; Oeemig JS; Kajander T; Iwai H *Nat Chem Biol* 2013, 9, 616. [PubMed: 23974115]
- (27). Dearden AK; Callahan B; Roey PV; Li Z; Kumar U; Belfort M; Nayak SK *Protein Sci* 2013, 22, 557. [PubMed: 23423655]
- (28). Poland BW; Xu MQ; Quirocho FA *J Biol Chem* 2000, 275, 16408. [PubMed: 10828056]
- (29). Xu MQ; Perler FB *EMBO J* 1996, 15, 5146. [PubMed: 8895558]
- (30). Chen L; Benner J; Perler FB *J Biol Chem* 2000, 275, 20431. [PubMed: 10770923]
- (31). Du Z; Zheng Y; Patterson M; Liu Y; Wang C *J Am Chem Soc* 2011, 133, 10275. [PubMed: 21604815]
- (32). Amitai G; Callahan BP; Stanger MJ; Belfort G; Belfort M *Proc Natl Acad Sci U S A* 2009, 106, 11005. [PubMed: 19541659]
- (33). Shah NH; Eryilmaz E; Cowburn D; Muir TW *J Am Chem Soc* 2013, 135, 5839. [PubMed: 23506399]
- (34). Cheriyan M; Pedomallu CS; Tori K; Perler F *J Biol Chem* 2013, 288, 6202. [PubMed: 23306197]
- (35). Stevens AJ; Sekar G; Shah NH; Mostafavi AZ; Cowburn D; Muir TW *Proc Natl Acad Sci U S A* 2017, 114, 8538. [PubMed: 28739907]
- (36). Ludwig C; Pfeiff M; Linne U; Mootz HD *Angew Chem Int Ed Engl* 2006, 45, 5218. [PubMed: 16823791]
- (37). Popp MW; Antos JM; Grotenbreg GM; Spooner E; Ploegh HL *Nat Chem Biol* 2007, 3, 707. [PubMed: 17891153]
- (38). Nguyen GK; Kam A; Loo S; Jansson AE; Pan LX; Tam JP *J Am Chem Soc* 2015, 137, 15398. [PubMed: 26633100]
- (39). Appleby-Tagoe JH; Thiel IV; Wang Y; Wang YF; Mootz HD; Liu XQ *Journal of Biological Chemistry* 2011, 286, 34440. [PubMed: 21832069]
- (40). Muir TW *Annu Rev Biochem* 2003, 72, 249. [PubMed: 12626339]
- (41). Ludwig C; Schwarzer D; Mootz HD *J Biol Chem* 2008, 283, 25264. [PubMed: 18625708]
- (42). Steipe B *Methods Enzymol* 2004, 388, 176. [PubMed: 15289071]
- (43). Gramespacher JA; Stevens AJ; Thompson RE; Muir TW *Protein Sci* 2018, 27, 614. [PubMed: 29226478]
- (44). Sun W; Yang J; Liu XQ *J Biol Chem* 2004, 279, 35281. [PubMed: 15194682]
- (45). Eisenberg D; Schwarz E; Komaromy M; Wall R *J Mol Biol* 1984, 179, 125. [PubMed: 6502707]

Figure 1: Protein splicing of canonical and atypical split inteins.

(a) Depiction of protein *trans*-splicing between an N-intein (Int^{N}) fused to an N-extein (Ex^{N}) and a C-intein (Int^{C}) fused to a C-extein (Ex^{C}). (b) Depiction of the canonical and atypical intein split sites. (c) Structural rendering of the canonical (red) and atypical (blue) split inteins corresponding to the split sites depicted in panel b and mapped onto the solution structure of the Npu-DnaE intein (PDB 2KEQ). The N- and C- termini are highlighted and labeled in black and green respectively.

Figure 2. Characterization of a consensus atypical (Cat) split intein.

(a) Pairwise sequence alignment of Cat and AceL* highlighting identical (black) and similar (gray) residues. (b) Reaction progress curve for Cat splicing at 30 °C. (c) Splicing rates for Cat and AceL* as a function of temperature (n = 3, error = SEM). AceL* is inactive at 50 °C. (d) Splicing rates for Cat and AceL* as a function of added Urea (n = 3, error = SEM). AceL* is not active in the presence of 2 M and 4 M Urea (NA).

Figure 3: Structural effects of Cat fragment association.

(a) ^1H - ^{15}N HSQC spectra of ^{15}N labeled Cat^{N} in free form (blue) and in complex with unlabeled Cat^{C} (red). (b) ^1H - ^{15}N HSQC spectra of ^{15}N labeled Cat^{C} in free form (blue) and in complex with unlabeled Cat^{N} (red). (c) Far UV circular dichroism spectra of Cat^{N} (black), Cat^{C} (blue) and the $\text{Cat}^{\text{N}} + \text{Cat}^{\text{C}}$ complex (red). (d) Size exclusion chromatograms of Cat^{N} (black), Cat^{C} (blue), and the $\text{Cat}^{\text{N}} + \text{Cat}^{\text{C}}$ complex (red).

Figure 4: Solution NMR structure of Cat.

(a) Backbone conformation of the 20 lowest energy conformers obtained in the structure calculation of the Cat^N (cyan) – Cat^C (brown) split intein complex. The N-terminus of Cat^N and C-terminus of Cat^C are colored in blue and red respectively, and the Cat^C solubility tag is rendered in transparent gray. Structures are shown with a 180° rotation (top and bottom renderings). (b) Cartoon depiction of the lowest energy conformer. Structures are shown with a 180° rotation (top and bottom renderings). (c) Zoom view of the Cat active site with Ala₁, Ser₇₅, His₇₈, and His₁₃₃ depicted as sticks. The distances between the carbonyl oxygen of Ala₁ and amide and hydroxyl protons of Ser₇₅ are indicated.

Figure 5. Localization of Disorder in the Cat Fragments.

(a) RP-HPLC chromatogram stack from the limited proteolysis of Cat^N (left), Cat^C (middle) and a 1:1 Cat^N + Cat^C complex (right) with samples quenched after the indicated times. (b) Sequence of Cat with the disordered regions of Cat^C highlighted in red and the protected center highlighted in beige. (c) Model of Cat disorder mapped onto the NMR structure with the N-intein highlighted in cyan, disordered region of Cat^C highlighted in red, and the protected center highlighted in beige. A zoom view of the active site is shown with the catalytic residues rendered as sticks.

Figure 6: Hydrophobic residues drive Cat association

(a) Surface rendering of Cat^N with hydrophobic residues colored in orange based on the normalized consensus hydrophobicity scale.⁴⁵ Cat^C is depicted as a cartoon (cyan). (b) Surface rendering of Cat^C with hydrophobic residues scaled in orange. Cat^N is depicted as a cartoon (cyan). (c) Equilibrium fluorescence anisotropy measurements of Fl-Cat^N (500 pM) in the presence of SUMO-Cat^C (indicated concentration) in low (100 mM NaCl, red) and high (500 mM NaCl, blue) salt buffers. (d) Concentration dependence of the observed rates of Fl-Cat^N+SUMO-Cat^C association in low (100mM NaCl, red) and high (500 mM NaCl, blue) salt buffers.

Figure 7: Extein Dependence of Cat.

(a) Schematic of the assay used to investigate the impact of local extein sequences on Cat splicing. An N-extein maltose binding protein (MBP) is fused to the Cat^N while a C-extein Green fluorescent protein (GFP) is fused to Cat^C. The native extein sequences (Phe₋₂, Glu₋₁, Cys₊₁, Glu₊₂, Phe₊₃) are shown within these fusion proteins. (b) Splicing rates for Cat in the presence of non-native C-extein residues (n = 3, error = SEM). Each indicated value corresponds to a single point mutation within the C-extein from the wild type (WT) sequence. (c) Splicing rates for Cat in the presence of non-native N-extein residues (n = 3, error = SEM). Each indicated value corresponds to a single point mutation within the N-extein from the wild type (WT) sequence. (d) Zoom view of the Cat active site with Cys₊₁, Glu₊₂, Asp₁₁₅, Asn₁₂₃, His₁₃₃, and Ala₁₃₄ depicted as sticks. (e) Zoom view of Cat active site with Glu₋₁, Ala₁, Ser₇₅, and His₇₈ depicted as sticks.

Table 1:

Statistics from NMR structure determination calculations of Cat complex in solution.

Parameter	Value
<i>Restraints</i>	
Distance restraints	3489
Unambiguous restraints	3283
Intra-residue	1667
Sequential	642
Short range	266
Long range	708
Ambiguous restraints	206
Dihedral angle restraints	180
<i>Structure statistics</i>	
NOE Violations > 0.5 Å	12 (+/- 4)
Dihedral violations > 5	0
Total Energy (kcal/mol)	-5074 (+/- 163)
<i>RMSD from mean structure</i>	
Backbone (all residues)	1.99 Å(+/- 0.4)
Heavy atoms (all residues)	2.52 Å(+/- 0.4)
Backbone (structured *)	1.19 Å(+/- 0.3)
Heavy atoms (structured *)	2.04 Å(+/- 0.3)
<i>Ramachandran plot analysis</i>	
Most favoured regions	85.7%
Additional allowed regions	13.5%
Generously allowed regions	0.8%
Disallowed regions	0.0%

* excluding exteins and solubility tag