Spontaneous Proton Transfer to a Conserved Intein Residue Determines On-Pathway Protein Splicing

Brian Pereira[1,2]†, **Philip T. Shemella**[3]†, **Gil Amitai**[1], **Georges Belfort**[2], **Saroj K. Nayak**[3]* and **Marlene Belfort**[1,4]*

[1]*Wadsworth Center, New York State Department of Health, Albany, NY 12201, USA*
[2]*Howard P. Isermann Department of Chemical and Biological Engineering, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
[3]*Department of Physics, Applied Physics, and Astronomy, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
[4]*Department of Biomedical Sciences, School of Public Health, State University of New York at Albany, Albany, NY 12201, USA*

The discovery of inteins, which are protein-splicing elements, has stimulated interest for various applications in chemical biology, bioseparations, drug delivery, and sensor development. However, for inteins to effectively contribute to these applications, an increased mechanistic understanding of cleavage and splicing reactions is required. While the multistep chemical reaction that leads to splicing is often explored and utilized, it is not clear how the intein navigates through the reaction space. The sequence of reaction steps must progress in concert in order to yield efficient splicing while minimizing off-pathway cleavage reactions. In this study, we demonstrate that formation of a previously identified branched intermediate is the critical step for determining splicing over cleavage products. By combining experimental assays and quantum mechanical simulations, we identify the electrostatic interactions that are important to the dynamics of the reaction steps. We illustrate, *via* an animated simulation trajectory, a proton transfer from the first C-terminal extein residue to a conserved aspartate, which synchronizes the multistep enzymatic reaction that is key to splicing. This work provides new insights into the complex interplay between critical active-site residues in the protein splicing mechanism, thereby facilitating biotechnological application while shedding light on multistep enzyme activity.

Corresponding authors. M. Belfort is to be contacted at Wadsworth Center, New York State Department of Health, Albany, NY 12201, USA. E-mail addresses: nayaks@rpi.edu; belfort@wadsworth.org.

† B.P. and P.T.S. contributed equally to this work.

Present addresses: B. Pereira, Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; P. T. Shemella, IBM Research-Zurich, Rüschlikon, 8803, Switzerland; G. Amitai, Molecular Genetics Department, Weizmann Institute of Science, Rehovot, 76100, Israel.

Introduction

An intein is an intervening sequence that is transcribed, translated, and subsequently removed at the protein level.[1] The intein is active in its own removal by catalyzing the process of protein splicing, a unique posttranslational phenomenon in which the flanking sequences, termed exteins, become joined together through a peptide bond as the intein is released. Protein splicing proceeds autocatalytically, without the need for cofactors; it requires only the intein and a few flanking extein residues.[1,2] These unusual chemical and biological properties of protein splicing predicate the utilization of inteins for a myriad of biotechnological applications, including protein purification, peptide cyclization, protein labeling, and biosensor development.[3-8] A challenge in engineering inteins is the complexity of protein splicing, which involves multiple coordinated reactions. Thus, in addition to enhancing fundamental understanding of the splicing mechanism, unraveling this complexity will be key to advancing such applications.

With a few exceptions,[9-11] protein splicing comprises four reactions in proceeding from the translation product, known as the precursor, to the mature gene product consisting of the ligated exteins.[12,13] In the first reaction, an N–X acyl shift (where X is S or O), the first residue of the intein, a cysteine or serine, attacks the preceding peptide carbonyl, forming a (thio)ester (Fig. 1a, step 1).[12,14,15] This intermediate maintains the linear order of the precursor sequence; thus, it is termed the linear (thio)ester or linear intermediate. The second step of protein splicing is a transesterification reaction in which the first residue of the C-extein, a cysteine, serine, or threonine, attacks the (thio)ester carbonyl and produces a (thio)ester between the N- and the C-extein (Fig. 1a, step 2).[16] This intermediate contains two N termini, belonging to the intein and N-extein, giving it a branched configuration; it is therefore termed the branched (thio)ester or branched intermediate.[12,16] In this state, the exteins are joined together but are still attached to the intein at its C terminus. Then, the last residue of the intein, an asparagine, cyclizes, resulting in irreversible cleavage of the succeeding peptide bond (Fig. 1a, step 3).[16-18] In the final step of protein splicing, the free exteins

undergo an X–N acyl shift, which reverts the (thio)ester to a peptide bond, yielding the mature gene product (Fig. 1a, step 4).[18]

The four steps of protein splicing may be classified as either (thio)ester-related or asparagine cyclization. These two activities are mechanistically independent and can serve as coordinates for the intein reaction space (Fig. 1a, coordinates R_N and R_C). Splicing is one pathway through this reaction space (Fig. 1a, steps 1–4); off-pathway reactions consist of single-cleavage events at either of the intein's termini. N-terminal cleavage occurs when an exogenous nucleophile, such as dithiothreitol or hydroxylamine, attacks the (thio)ester bonds of the linear and branched intermediates. Because the rate of N-terminal cleavage of the branched intermediate is significantly greater than that of the linear intermediate (discussed below), we only show the former in the intein reaction space (Fig. 1a, step N). C-terminal cleavage occurs when asparagine cyclization precedes formation of the branched intermediate (Fig. 1a, steps C). To achieve splicing and avoid these off-pathway cleavage events, the intein must navigate the reaction space by the modulation and coordination of competing reactions; it is not well understood how the intein completes this. Additionally, the roles of some highly conserved residues have yet to be firmly established. For these reasons, we have investigated the protein splicing mechanism with a focus on the role of a well-conserved active-site residue, which we had previously identified as a determinant of splicing over cleavage[19,20] and, when mutated, had found utility in promoting cleavage required for protein purification.[20]

Our study utilizes the *Mycobacterium tuberculosis* (Mtu) RecA intein because of its potential as an antimycobacterial target[21,22] and because of our knowledge on this intein's structure and function.[19,20,23-25] On the basis of an initial directed evolution study and subsequent analyses, residue Asp422 was identified as playing an important role in protein splicing.[19,20] This residue is located within a region conserved among inteins, termed Block F (also termed the C2 motif).[26,27] As the fourth position of Block F, the residue is well conserved as an aspartate; out of the 552 reported inteins (as of 10/21/10), aspartate occurs in 324 of them.[28] In two of these inteins, the equivalent aspartate residue has been

---

Fig. 1. Splicing and cleavage reactions of the Mtu RecA intein. (a) The general reaction mechanism of protein splicing. The scheme is depicted along pseudo reaction coordinates corresponding to the activity: R_N is (thio)ester-related and R_C is asparagine cyclization-related. The N-extein, intein, and C-extein are colored blue, red, and green, respectively. For the Mtu RecA 110Δ383 mini-intein, the three mechanistic residues are Cys1 for step 1, Cys+1 for step 2, and Asn440 for step 3. Arrows depicting the four steps of splicing are on a yellow background. The reaction described by QM is boxed. (b) The Mtu RecA mini-intein model used for QM simulations based on the intein crystal structure.[19] The intein's terminal residues (Cys1 and Asn440), the highly conserved Block-F aspartate (Asp422), and the first residue of the C-extein (Cys+1) are shown as sticks. The residue numbers are those of the full-length intein.
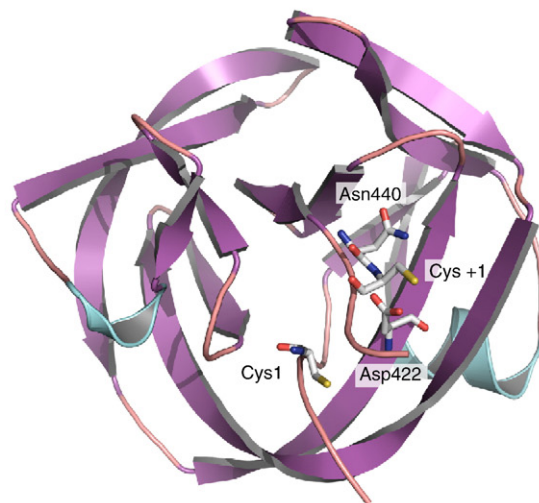
(a)
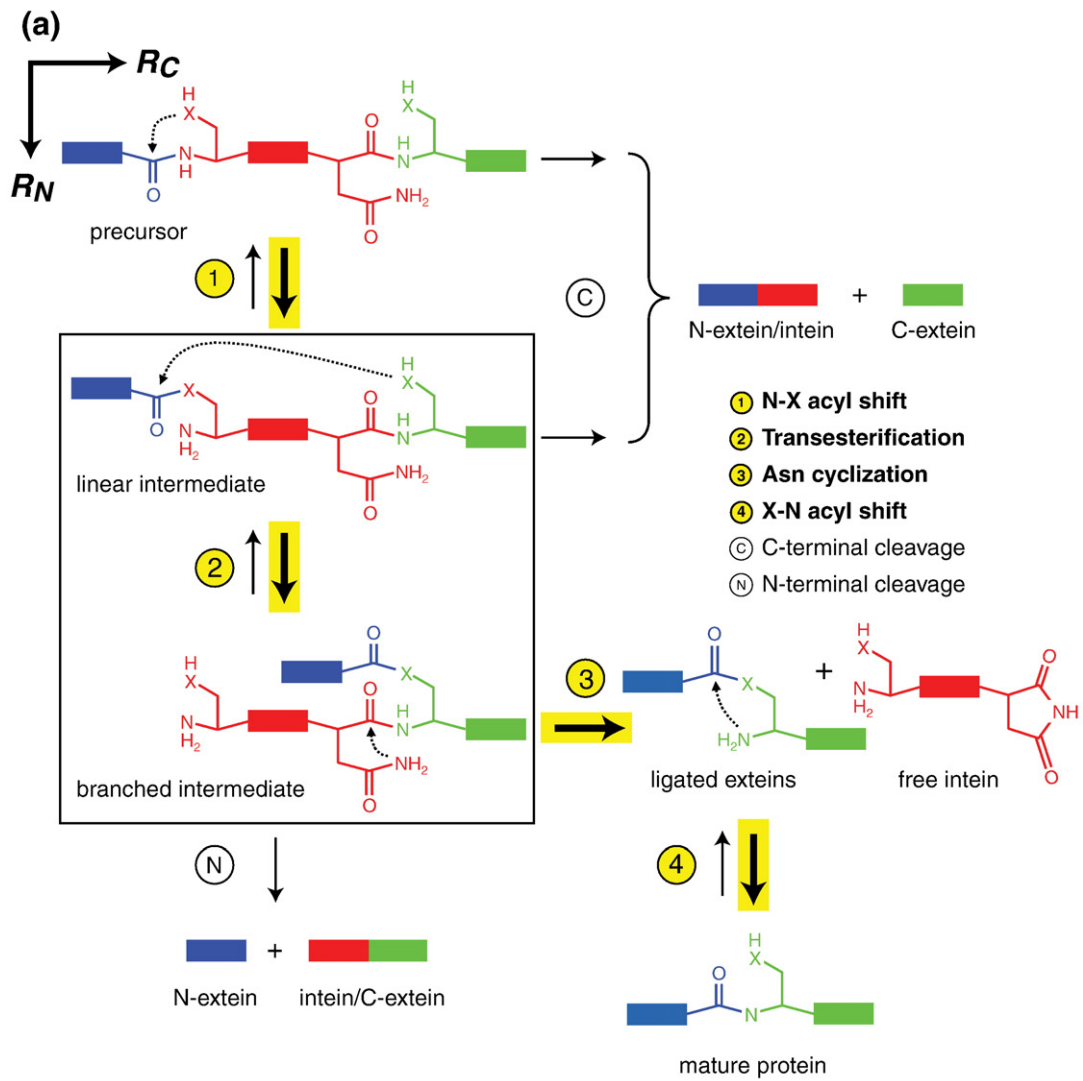
(b)

Fig. 1 (*legend on previous page*)

shown to have a role in intein activity.[29,30] With regard to tertiary structure, the Block-F aspartate lies within the active site between the intein's N and C termini, where it may be capable of interacting with either terminus or possibly both (Fig. 1b).[19,29,30] By combining experimental assays, the existing crystal structure, and quantum mechanical [quantum mechanics (QM)] simulations, we identify the role of Asp422, which is further illustrated in an animated simulation trajectory. This acidic amino acid accepts a proton from the first residue of the C-extein after linear thioester formation, thereby catalyzing the transesterification reaction. This important role gives insight into how the intein navigates the reaction space to achieve on-pathway splicing, knowledge that is of use not only in manipulating splicing levels for application but also for understanding the mechanism of other proteases that have mechanistic parallels.

Results

An acidic residue at position 422 is sufficient for splicing

To evaluate *in vivo* intein activity, we utilized a tripartite fusion protein consisting of Mtu RecA mini-intein (I) variants with maltose-binding protein (M) and the C-terminal domain of I-*Tev*I (C) as the N-extein and C-extein, respectively. The fusion protein (MIC), with the 168-amino-acid 110Δ383 mini-intein,[23] was expressed in *Escherichia coli*, and the whole-cell lysate was separated *via* SDS-PAGE and visualized by Coomassie staining. Bands corresponding to MIC-related products were identified by molecular mass and confirmed through Western blotting (Fig. 2a). For the intein with wild-type Asp422 (D), we observed complete disappearance of the precursor (MIC, 74.7 kDa) and a significant amount of ligated exteins (MC, 56.1 kDa) and free intein (I, 18.6 kDa), verifying that the Mtu RecA intein mediated protein splicing in this context (Fig. 2a, lanes 1). There were also some free N-exteins, (M, 43.0 kDa), which are splicing side-products; their presence indicates that splicing efficiency was <100%, which is not uncommon for inteins in heterologous contexts.[1]

A previously isolated intein variant containing a D422G mutation exhibited different intein activity.[20] This variant yielded no spliced product but a significant amount of fused N-extein–intein (MI, 61.5 kDa; Fig. 2a, lanes 6) and C-extein (data not shown), which are other splicing side-products. It is clear that the D422G mutation disrupted at least one of the reactions of protein splicing. However, it was uncertain whether the disruption was due to the loss of the aspartate or the gain of glycine at position 422. To investigate the basis of the phenotype, we generated and assayed intein variants with Asp422 mutated to glutamate (E), leucine (L), asparagine (N), or alanine (A), amino acids with diverse side-chain properties. Of these variants, only the intein with the D422E mutation generated any spliced product, albeit to a lesser extent than the wild type (Fig. 2a, cf. lanes 1 and 2 with lanes 3–6). Asn422 and Ala422 produced only MI, exhibiting intein activity similar to that of Gly422, and Leu422 was inactive over this timescale. These results show that an acidic amino acid at position 422 is sufficient and possibly necessary for the Mtu RecA intein to perform protein splicing.

Formation of branched thioester intermediate

Mutations to the intein that prevent protein splicing act by impairing some combination of the reactions shown in Fig. 1a. To deconvolute a mutation's influence into step-specific effects, we introduced an N440A mutation into the Mtu RecA intein variants described above. This mutation precludes the asparagine cyclization reaction (Fig. 1a, step 3); however, the intein should still be capable of forming the branched thioester intermediate. Therefore, we limited our investigation to the first two reactions of splicing (Fig. 1a, steps 1 and 2) and again observed products by Western blotting of SDS-PAGE gels. Previously, the branched intermediate has been observed in SDS-PAGE as a band that migrates more slowly than the precursor due to its nonlinear configuration.[12] Similarly, for the Mtu RecA intein with Asp422, we discerned a slowly migrating band, and through Western blotting, we confirmed that this band contained M, I, and C (Fig. 2b, lanes 2, and Fig. S1). Furthermore, the band was absent when Cys+1 was mutated to an alanine (Fig. 2b, lanes 1), as expected because Cys+1 is required for formation of the branched intermediate (Fig. 1a, step 2). The branched intermediate, reported for the first time for the Mtu RecA intein, disappeared with mutation of Asp422 to glutamate, asparagine, alanine, or glycine (Fig. 2b, lanes 3–6).

We further examined branched intermediate formation with a recently developed fluorescence resonance energy transfer (FRET)-based fusion protein, which provides for a more robust, kinetic comparison of intein variants.[2] The fusion protein consists of the intein with FRET-active cyan fluorescent protein and yellow fluorescent protein as the N-extein and C-extein, respectively (Fig. 3a). In the intein precursor state, the fusion protein exhibits FRET; this FRET is lost when one fluorescent protein is separated from the other, such as with N-terminal
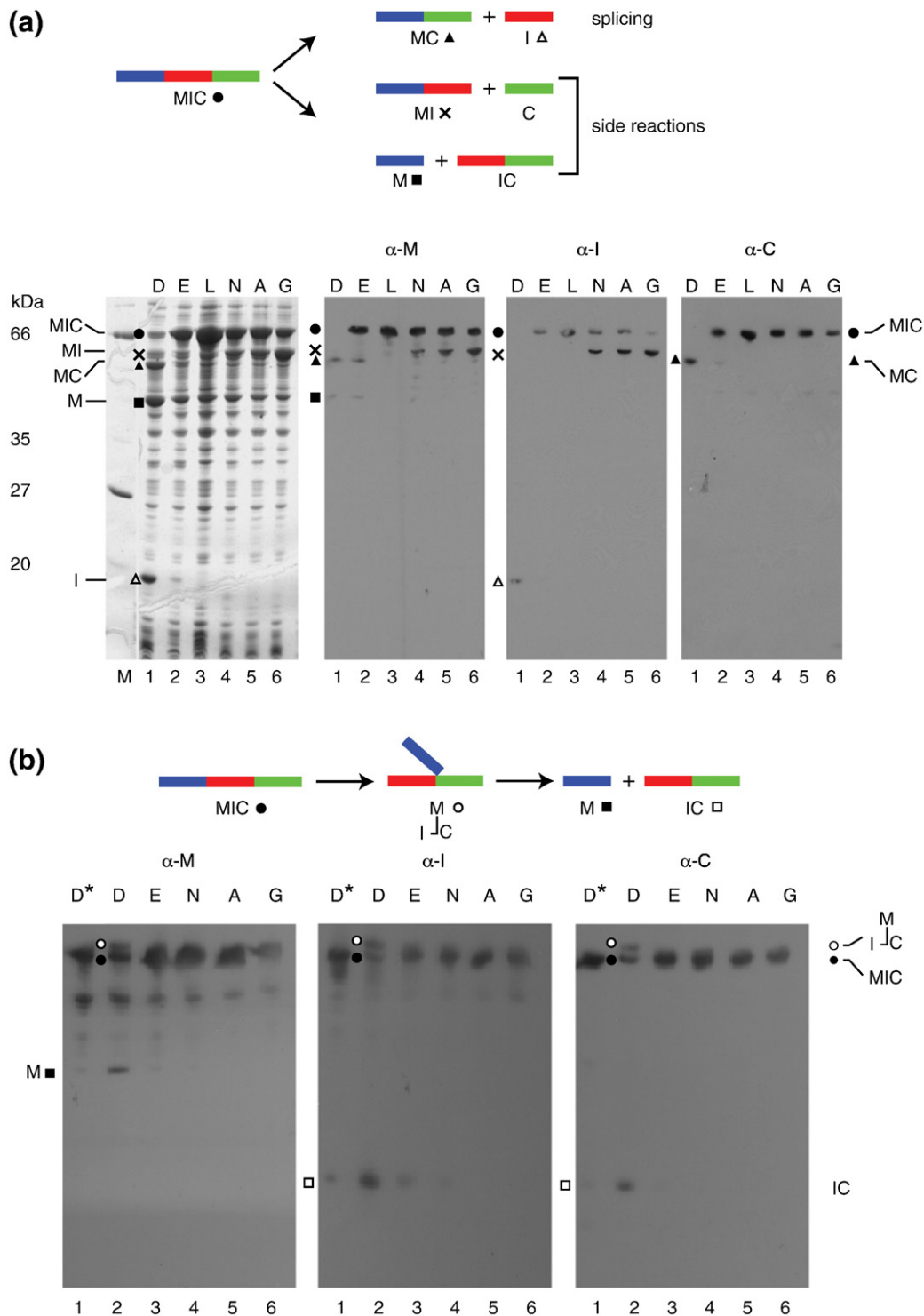
Fig. 2. Reaction products with Asp422 variants in the MIC context. (a) Reaction products with splicing-competent intein. In the schematic, the MIC precursor and products are depicted as in Fig. 1a. SDS-PAGE and Western blots of lysates with intein variants are shown below. The amino acid at position 422 is above each lane (single-letter code). The antibodies (α) used for each blot, MIC-related bands, and molecular weight standards are labeled accordingly. It must be noted that lysate background has some overlap with product bands in the Coomassie-stained gels. (b) Reaction products with Asp422 variants in an N440A background. Western blots of lysates with intein variants containing the N440A mutation are shown, labeled as in (a). Unlabeled bands do not correspond to any known product and are likely due to proteolysis. The branched intermediate is marked with an open circle. D∗ represents Asp422 with a C+1A mutation.
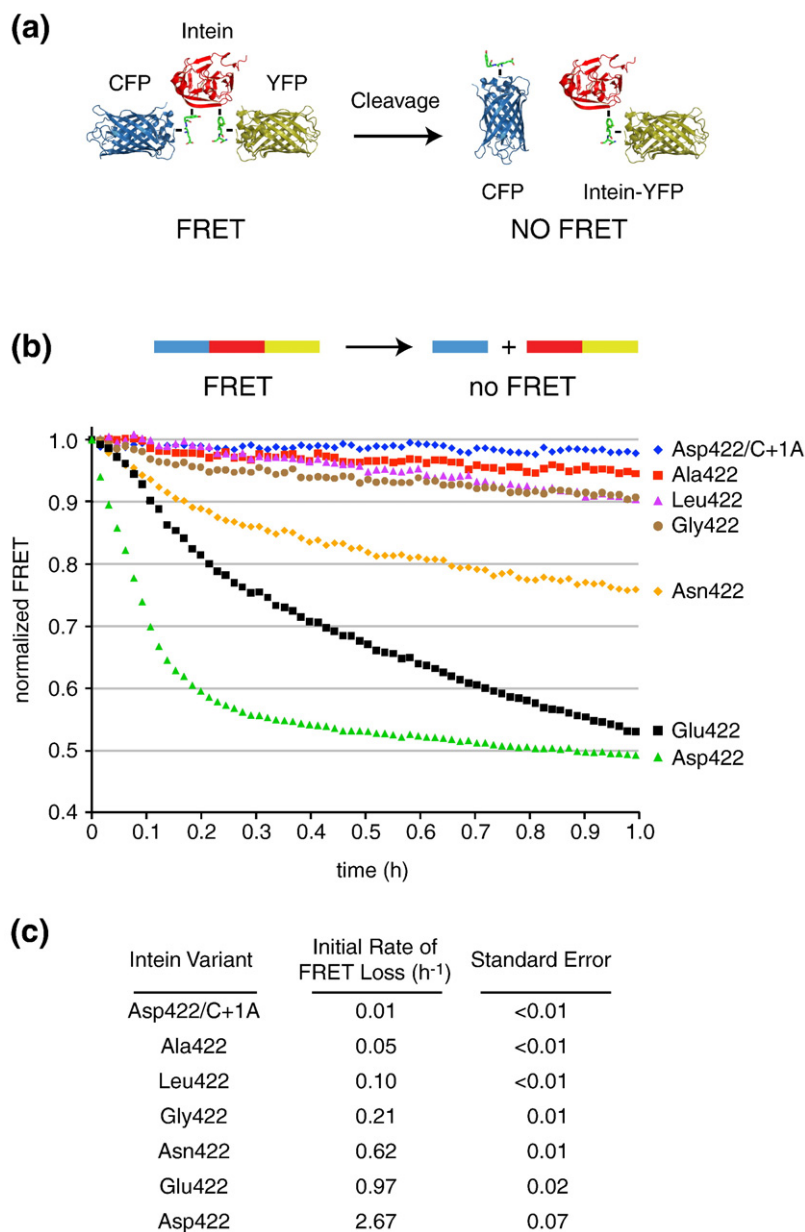
Intein Variant	Initial Rate of FRET Loss (h⁻¹)	Standard Error
Asp422/C+1A	0.01	<0.01
Ala422	0.05	<0.01
Leu422	0.10	<0.01
Gly422	0.21	0.01
Asn422	0.62	0.01
Glu422	0.97	0.02
Asp422	2.67	0.07

Fig. 3. FRET-based N-extein cleavage assay for Asp422 variants. (a) FRET reporter. The precursor (C-I-Y) is FRET active, with N-terminal cleavage of N440A constructs being associated with loss of FRET. (b) Loss of FRET over time for intein variants. FRET values have been normalized by the initial FRET values. The amino acid at position 422 is labeled, with Asp422 being wild type. The fusion constructs all have an N440A mutation to prevent C-terminal cleavage. (c) Initial rates of FRET loss for intein variants.

cleavage (Fig. 1a, step N). As stated above, N-terminal cleavage is the reaction of exogenous nucleophile with the thioester intermediates; therefore, for intein variants in the presence of excess nucleophile, the loss of FRET is indicative of the formation of thioester intermediates.

We cloned the intein variants with the N440A mutation into the FRET fusion protein, expressed the fusion proteins, added excess hydroxylamine to induce N-terminal cleavage, and measured FRET over time (Fig. 3b). While the wild-type Asp422 exhibited an immediate, rapid decline in FRET, there was negligible change in FRET for Asp422/C+1A (Fig. 3b and c). This disparity confirms that N-terminal cleavage is due mainly to nucleophilic attack of the branched thioester, as observed for the *Synechocystis* sp. PCC6803 DnaE intein.[31] On this basis, the rates of FRET loss reveal that mutation of Asp422 considerably reduces branched thioester formation (Fig. 3b and c); furthermore, FRET analysis exposes the differences between these mutants. The Ala422, Gly422, and Leu422 variants are all >10-fold less active than the wild-type Asp422, whereas Asn422 and Glu422 are only 4.3- and 2.8-fold less active (Fig. 3c). These rates can explain why Asp422 and Glu422 yield spliced product (Fig. 2a, lanes 1 and 2), while the other mutants do not (Fig. 2a, lanes 4–6).

QM analysis suggests that Asp422 accepts a proton to catalyze transesterification

To explore intein activity at the atomic level and to further elucidate the role of Asp422, we conducted QM simulations of the intein active site. The hybrid QM and molecular mechanics (MM) method has been an important tool for understanding biochemical processes,[32-34] and similar computational schemes have been used with success.[35-39] Our study consisted of simulating protein splicing by computationally progressing through the reaction pathway. Various chemical states along the pathway, including the linear and branched intermediates, were relaxed using QM/MM simulations, and the energies of these states were calculated. Results for the N–S acyl shift reaction (Fig. 1a, step 1) and asparagine cyclization reaction (Fig. 1a, step 3) have been described previously.[40,41] Here, we present the findings relevant to the transesterification reaction (Fig. 1a, step 2), beginning with the atomic level structure of the linear thioester intermediate.

In protein splicing reaction schemes, the linear intermediate is typically depicted with a $-NH_2$ group at the intein's N terminus (Figs. 1a and 4a).

However, based on conventional pK_a values of solvated protein N termini ($\sim 8–9$[42]), the amino group is likely to become protonated, resulting in a $-NH_3^+$ group. As such, in our computational study, we added a proton to the linear intermediate (Fig. 4b). Upon relaxation of this chemical state with QM simulations, a proton spontaneously transferred from Cys+1 to Asp422 due to the repulsive electrostatic interaction between the N-terminal $-NH_3^+$ and the Cys+1 $-SH$ groups. This barrierless proton transfer, shown as a movie in the Supplementary Data, yielded an energetically more stable structure, by ~ 2 kcal/mol. More importantly, the resulting deprotonated Cys+1 side chain was a negatively charged thiolate group, primed for nucleophilic attack; effectively, Cys+1 became activated for transesterification and formation of the branched intermediate (Fig. 4d).

The importance of the active-site interactions to this proton transfer was analyzed by computationally mutating certain residues and conducting similar QM/MM simulations with these intein mutants. With one mutant, in which Cys+1 was replaced in the simulation with serine, proton transfer did not occur spontaneously; instead,
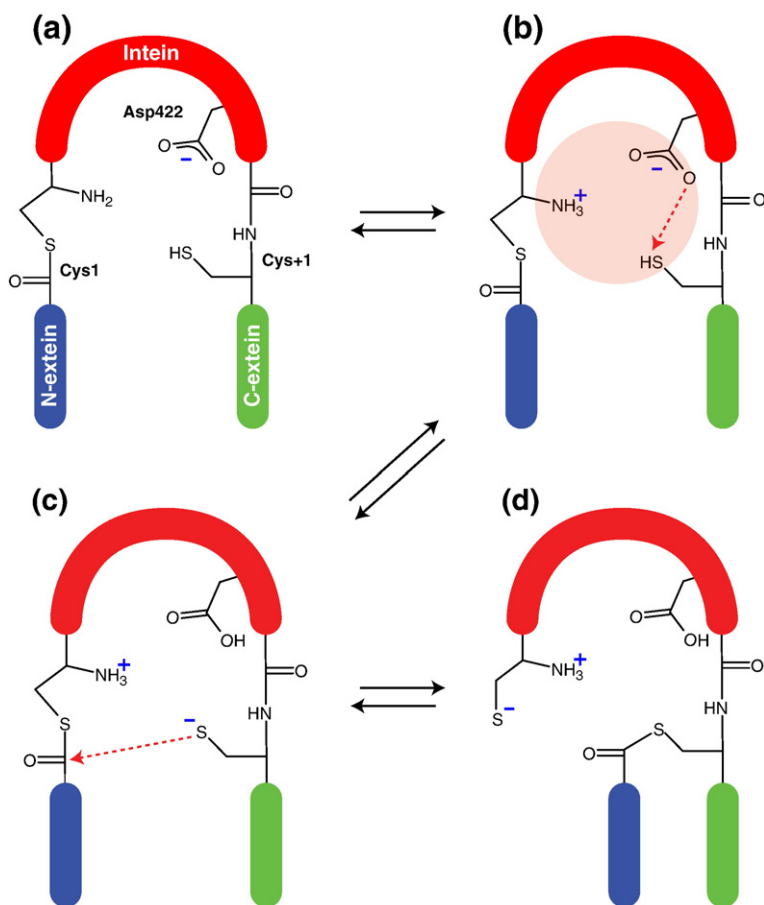

Fig. 4. The reaction from the linear thioester to the branched thioester (Fig. 1a, step 2). (a) Linear thioester intermediate as shown in the general reaction mechanism. (b) A proton was added to the N terminus to mimic the state as expected by typical pK_a values.[42] The electrostatic repulsion between the N-terminal $-NH_3^+$ and the Cys +1 $-SH$ groups causes Cys +1 to transfer a proton to Asp422, thus electrostatically stabilizing the region. Red broken lines show the path of attacking electrons. (c) As a thiolate, Cys+1 is primed for attack on the linear thioester bond. (d) The branched thioester intermediate is produced *via* transesterification once Cys+1 is ionized.

manually simulating the transfer of a proton from Ser+1 to Asp422 required 12.2 kcal/mol. This energy, compared with that for the wild-type Cys +1 that was slightly exothermic, is consistent with the difference in pK_a values of serine and cysteine side chains ($\sim$13 and 8.3, respectively[43]). The hydroxyl group on the serine side chain is extremely polar, whereas the thiol of the cysteine side chain is nonpolar. Reflected by its exceedingly high typical pK_a value, the Ser+1 side chain better accommodated the charge on the linear intermediate's N terminus ($-NH_3^+$). The nonpolar Cys+1 side chain experienced energetic destabilization that caused the thiol group's pK_a to lower and Cys+1 to donate its proton to Asp422. Though serine and threonine at +1 are expected to decrease splicing for the Mtu RecA intein on the basis of these presented interactions, inteins with native Ser+1 or Thr+1 may require additional interactions.

In another simulation, Asp422 was mutated to asparagine, a polar residue roughly the same size as aspartate. Transfer of the proton from Cys+1 to the side-chain oxygen of Asn422 required 45 kcal/mol, a substantial energy hurdle for the deprotonation of Cys+1. We also mutated Asp422 to glycine, which is unable to act as a proton acceptor. During the simulation with this mutant, the proton from Cys+1 did not transfer to any residue. Without stabilization from the ionized Cys+1, the N-terminal $-NH_3^+$, due to the changes in the active-site interactions and active-site architecture, gave a proton to a nearby histidine residue (His73). We previously proposed that this histidine residue plays a role in the N–S acyl shift,[41] highlighting the dynamic balance of active-site electrostatics throughout the different steps of protein splicing. The role of Asp422 in this balance is to accept a proton from Cys+1 after linear thioester formation, thereby catalyzing transesterification.

Discussion

In this work, we have combined molecular biology and QM to investigate the reaction space of protein splicing and the role of a conserved aspartate, Asp422, in balancing the electrostatics at the active site toward splicing. The results are important for establishing how spontaneous proton transfer to Asp422 determines on-pathway protein splicing over off-pathway cleavage events. Protons are known as the most important effectors in enzyme catalysis,[44] and this journey along the splicing pathway is achieved by Asp422 accepting a proton from the Cys+1 thiol group, thereby catalyzing the transesterification reaction and linking the N- and C-terminal reactions.

Though N- or C-terminal cleavage can be easily isolated through mutation of critical residues,[13] it is a challenge to increase splicing without consequently stimulating the cleavage side reactions. Furthermore, splicing-based applications typically require modulated activity. Intein-based biosensors, for example, must be designed to have little initial splicing and subsequent rapid reactivity upon detection of the analyte.[45-47] Therefore, engineering inteins for splicing applications would benefit from an understanding of the intricate balance of factors, especially the active-site electrostatics, that direct the intein through the reaction space.

This progression through the intein reaction space can be represented along two reaction coordinates: one based on thioester-related activity and one based on asparagine cyclization (Fig. 1a, coordinates R_N and R_C, respectively). The two-dimensional reaction space described by these coordinates includes the pathways for splicing, N-terminal cleavage, and C-terminal cleavage. It is helpful to note the competing reactions that differentiate the three pathways. One set of competing reactions determines the fate of the branched intermediate: exogenous nucleophilic attack on the thioester bond results in N-terminal cleavage, and asparagine cyclization leads to splicing (Fig. 1a, steps N and 3, respectively). This point of divergence suggests that, for greater splicing yield, the rate of asparagine cyclization should be maximized relative to the rate of exogenous nucleophilic attack.

However, we must consider the remaining reaction space. Asparagine cyclization is also a competing reaction with the N–S acyl shift and transesterification combination (Fig. 1a, steps 1 and 2); if asparagine cyclization precedes formation of the branched intermediate, C-terminal cleavage occurs (Fig. 1a, step C). On this basis, more splicing will occur when asparagine cyclization is minimized relative to branched intermediate formation or, in other words, when branched intermediate formation is maximized. In a new study, it was shown that the rate of asparagine cyclization is not constant along the R_N coordinate; rather, the reaction proceeds more rapidly for the branched intermediate than for the precursor or linear intermediate[48] (i.e., step 3 is faster than step C in Fig. 1a). The observed effect is due to a change in local structure upon branched intermediate formation. In our work, we observed a range of C-terminal cleavage for Leu422, Asn422, Ala422, and Gly422, but these mutants did not splice (Fig. 2a, lanes 3–6). This result suggests that the critical aspect for splicing is the rate of branched intermediate formation.

Examination of branched intermediate formation revealed a balance of electrostatic interactions critical to transesterification. One aspect of this balance is the acidic side chain of Asp422. Simulation results show that Asp422 accepts a proton from Cys+1. This proton transfer activates Cys+1 for nucleophilic attack on the linear thioester; the protonated Asp422 may then use a hydrogen bond

to stabilize transesterification transition states or may even act as an eventual proton donor. In both cases, the proton transfer from Cys+1 to Asp422 effectively catalyzes the transesterification reaction, and correspondingly, the intein exhibits splicing and branched intermediate formation (Fig. 2a and b). The other important component is Cys+1. Cys+1 was found to have a pK_a value of 5.8, more than 2 units lower than typical cysteines;[49] this may be indicative of its interaction with Asp422. Asp422 may play a related, but different, mechanistic role in inteins with native Ser+1 or Thr+1 owing to the difficulty in removing a proton from these residues.

Experimental mutation of the aspartate to the acidic glutamate also enables splicing to proceed, but splicing is reduced by the mutation. It is probable that the active site is ideal for aspartate's geometry, not the larger glutamate. Glutamate was not studied with QM due to significant disruption of the structural model, and interestingly, in crystal structures, a glutamate at residue 422 is more disordered than an aspartate (Patrick Van Roey, personal communication). Mutations of residue 422 to leucine, asparagine, alanine, and glycine prevented splicing; none of these amino acids can effectively act as proton acceptors. Beyond the Mtu RecA intein, a significant number of inteins have a polar residue at the equivalent Block-F position (142 inteins have a cysteine, serine, or threonine).[28] These polar residues are capable of proton transfer and may play a role similar to aspartate, or the inteins with these residues may utilize a different mechanism as revealed by recent work.[11]

A previous study utilizing site-directed mutagenesis showed that the equivalent aspartate plays a role in N-terminal cleavage.[30] Here, for the *Methanococcus jannaschii* KlbA intein, which splices through an alternative mechanism, N-terminal cleavage corresponds to formation of the branched intermediate. In a model based on the NMR structure of the *M. jannaschii* KlbA intein, the Block-F aspartate is within hydrogen-bonding distance to the Cys+1 side chain. On the basis of this model, the authors postulated that the aspartate plays a role in deprotonation of Cys+1. Our work expands on this by revealing the deprotonation with QM simulation.

An interesting feature of this Block-F residue is that it may possibly play multiple roles in intein activity. For example, the equivalent aspartate in the *Synechocystis* sp. PCC6803 DnaB intein was proposed to help stabilize an oxyanion state during asparagine cyclization.[29] Correspondingly, mutagenesis of the Block-F aspartate has been shown to affect asparagine cyclization.[19,20] Furthermore, in a new class of inteins, the equivalent Block-F residue is a cysteine, and it forms a different branched intermediate.[11] All these studies highlight the importance and multifunctionality of this Block-F residue.

There may be mechanistic parallels to other proteases in nature, extending this study beyond inteins. For example, various proteases are expressed as inactive zymogens, which are activated by proteolysis.[50-53] Among them are blood coagulation proteins such as prothrombin and factor-XIa, which must be processed to prevent bleeding disorders,[50,53] and kallikrein and neuropsin, which are useful as cancer markers.[51,52] These zymogens are activated once their N-terminal regions are proteolytically cleaved at a specific position. These proteases sense the cleavage event by formation of a salt bridge between the newly formed N-terminal amine and the carboxyl oxygen of a cardinal aspartate residue juxtaposed to the catalytic serine within the active site. This arrangement resembles that which accommodates the intein proton transfer described in our current work. The parallel in the intein reaction pathway is a thioesterification event that subsequently leads to autocleavage, producing a free primary amine group that interacts with the carboxyl oxygen of Asp422. It will be interesting to determine, by exploring activity at the atomic level, if such a proton transfer applies to similar processes in the activation of these other proteases.

Conclusions

Inteins have already contributed to and have future potential application in chemical biology, bioseparations, biomedicine, drug delivery, sensor development, and basic research. It is important for these and other applications of inteins that their use be versatile, robust, inexpensive, and easy to perform when used to produce the desired chemical modification. Thus, inteins need to splice faster, more efficiently, and under different solution conditions and possibly be triggered by other mechanisms than just temperature and pH. Currently, there are several developing biotechnologies that utilize protein splicing, such as purification of cytotoxic proteins,[54] sensor development,[45,46] and protein labeling.[5] For these and similar applications, it is important to reduce the off-pathway reactions: N- and C-terminal cleavage. In this study, we have combined biochemistry, genetics, and QM modeling to describe the intein reaction space and the factors involved in determining the reaction outcome. We have shown that splicing correlates with formation of the branched intermediate and that a conserved aspartate residue plays a significant role in this reaction. We have also identified a series of critical electrostatic interactions, induced by the formation of a free amino terminus, that control the dynamics of the reaction steps. This result consisted of an inherent proton-based regulation yielding synchronization and trafficking of a multistep enzymatic

reaction. These insights will provide a sound basis for engineering inteins for splicing applications. Our multidisciplinary approach and this particular case may also give broader insight into enzymatic processes carried out *via* multiple steps.

Materials and Methods

Plasmids and intein variants

The intein used in this study is the 168-residue engineered Mtu RecA mini-intein, 110Δ383;[23] it consists of residues 1–110 and 383–440, where residue numbers correspond to the full-length 440-amino-acid intein.[55] N-extein residues are numbered $-n, \ldots, -2, -1$, and C-extein residues are numbered $+1, +2, \ldots, +n$. For study of *in vivo* intein activity, we utilized the pMIC plasmid, which encodes the tripartite fusion protein MIC. MIC consists of *E. coli* maltose-binding protein (M), Mtu RecA intein (I), and the C-terminal domain of the homing endonuclease I-*Tev*I (C) from the bacteriophage T4.[23,56] Mutants of the engineered 110Δ383 Mtu RecA mini-intein were previously cloned into pMIC.[19,20] One of these mutants, an active splicing mutant, contains a single V67L mutation, with residue 422 maintained as the wild-type aspartate. The other mutant, a rapidly cleaving mutant, contains D24G (phenotypically silent), V67L, and D422G mutations.[20] In this study, these two mutants are referred to as "Asp422" and "Gly422." Other intein variants, designed on the basis of varying the side-chain properties of residue 422, were generated by site-directed mutagenesis of the Asp422 and Gly422 variants.

In vitro intein activity was investigated through the use of a plasmid (pCIY) that encodes a FRET-active fusion protein, cyan fluorescent protein–intein–yellow fluorescent protein.[2] The Mtu RecA 110Δ383 mini-intein variants and five N-extein and three C-extein residues were cloned from pMIC into the FRET plasmid.

SDS-PAGE and Western blotting

The intein-containing plasmids, pMIC, were transformed into *E. coli* JM109 and grown to $OD_{600} = 0.4–0.6$ at 37 °C. We induced expression of MIC precursor by adding 1 mM isopropyl-β,D-thiogalactopyranoside and incubating the cultures at 37 °C for 3 h. After induction, cells were collected by centrifugation, resuspended in lysis buffer (50 mM Tris, pH 8.0, and 2 mM ethylenediaminetetraacetic acid), and lysed by sonication. The soluble fractions of the cell lysates were mixed with 4× sample loading buffer [2% SDS, 10% glycerol, 0.0625 M Tris (pH 6.8), and 0.01% bromophenol blue], with or without 5% β-mercaptoethanol at a ratio of 3:1. The samples were heated at 95 °C for 3 min, and the proteins were resolved by SDS-PAGE. Some gels were visualized with Coomassie blue stain, while the proteins in other gels were transferred to PVDF membranes (Bio-Rad) for Western blotting. The membranes were incubated with antibodies to (α-) maltose-binding protein, Mtu RecA intein, and I-*Tev*I. The membranes were washed to remove any nonspecific binding and then incubated with either α-mouse immu-

noglobulin G or α-rabbit immunoglobulin G, depending on the primary antibody. These antibodies were conjugated with horseradish peroxidase, and upon mixture with the detection mix (GE Healthcare), they resulted in emitted light and, subsequently, exposure of X-ray film. Comparison with protein markers allowed identification of each MIC-related band.

Fluorescence resonance energy transfer

Intein variants in the context of the FRET plasmid were transformed into *E. coli* MC1061. Cultures were grown to $OD_{600} = 0.4–0.6$ at 37 °C, and protein expression was induced with 0.2% arabinose at 25 °C for 6 h. Cells were harvested and lysed in bacterial protein extraction reagent (B-PerII; Thermo Scientific) by vortexing for 1 min, and the lysates were clarified over Ni-NTA spin columns (Qiagen). The column eluates were mixed 1:8 with FRET buffer (20 mM Tris, pH 8.0, and 10 mM ethylenediaminetetraacetic acid) and incubated at 37 °C for 4 h to allow for equilibration. They were then transferred to 96-well plates (black), and hydroxylamine in FRET buffer was added to 100 mM.[2] The wells were excited at 400 nm, and emissions were measured at 485 nm and 540 nm over time. FRET was calculated as the emission ratios of 540 nm to 485 nm and then normalized by the initial FRET values. The averages of triplicate samples are presented, while initial rates and standard errors were determined through simple linear regression.

Molecular simulations

A QM/MM simulation scheme was used to compute atomic structure, energetics, and electronic properties of systems of interest.[57-59] The structure was based on the Mtu RecA crystal structure (Protein Data Bank code 2IN0).[19] Short peptide sequences based on the native extein residues were computationally appended,[40] and equilibration with full classical molecular dynamics was performed prior to the QM/MM investigation. The entire system consisted of 6314 atoms: 2351 belonging to the protein and 3693 to 1231 explicit water molecules. The QM active site includes 130+ atoms and is based on the following regions: the N-terminal active site including the N-extein residue Lys-1 and the intein residues Cys1 and Leu2; conserved residues Thr70, His73, and Asp422; and the C-terminal active site including Val438, His439, Asn440, Cys+1, and Ser+2. Nine water molecules were included explicitly in the QM calculations. For the QM region, we used first-principles density functional methods, specifically the B3LYP functional[59] with the 6-31G(d, p) basis set.[60] For the classical MM region, the AMBER force field[61] was used. The accuracy of the functional and basis sets used here was tested and found to provide satisfactory results, and similar methodology has been used previously.[32,40] The size of the quantum region considered here was further increased with identical results, therefore validating our computational conclusions. We have not included dynamic effects of structural changes, which may influence quantitative estimates for energy barriers. However, since our primary goal was to study relative energy differences between various mutants of similar size, dynamic effects should not alter

our overall conclusions.[62] Considering other configurations of Asp422 and then finding further functions are exceedingly complex and computationally difficult. Here, we present one function for Asp422 that is based on one configuration. This role is chemically logical and is consistent with results in literature as well as our own experimental effort.

Supplementary materials related to this article can be found online at doi:10.1016/j.jmb.2010.12.024

References

1. Perler, F. B., Davis, E. O., Dean, G. E., Gimble, F. S., Jack, W. E., Neff, N. *et al.* (1994). Protein splicing elements: inteins and exteins—a definition of terms and recommended nomenclature. *Nucleic Acids Res.* **22**, 1125–1127.
2. Amitai, G., Callahan, B. P., Stanger, M. J., Belfort, G. & Belfort, M. (2009). Modulation of intein activity by its neighboring extein substrates. *Proc. Natl Acad. Sci. USA*, **106**, 11005–11010.
3. Yamazaki, T., Otomo, T., Oda, N., Kyogoku, Y., Uegaki, K., Ito, N. *et al.* (1998). Segmental isotope labeling for protein NMR using peptide splicing. *J. Am. Chem. Soc.* **120**, 5591–5592.
4. Belfort, M., Derbyshire, V., Stoddard, B. & Wood, D. (2005). In *Inteins and Homing Endonucleases* (Belfort, M., Derbyshire, V., Derbyshire, B., & Wood, D., eds.), Springer-Verlag: Berlin Germany.
5. Hahn, M. E. & Muir, T. W. (2005). Manipulating proteins with chemistry: a cross-section of chemical biology. *Trends Biochem. Sci.* **30**, 26–34.
6. Muir, T. W. (2006). Cutting out the middle man. *Nature*, **442**, 517–518.
7. Ozawa, T. (2006). Designing split reporter proteins for analytical tools. *Anal. Chim. Acta*, **556**, 58–68.
8. Vila-Perello, M. & Muir, T. W. (2010). Biological applications of protein splicing. *Cell*, **143**, 191–200.
9. Southworth, M. W., Benner, J. & Perler, F. B. (2000). An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile. *EMBO J.* **19**, 5019–5026.
10. Amitai, G., Dassa, B. & Pietrokovski, S. (2004). Protein splicing of inteins with atypical glutamine and aspartate C-terminal residues. *J. Biol. Chem.* **279**, 3121–3131.
11. Tori, K., Dassa, B., Johnson, M. A., Southworth, M. W., Brace, L. E., Ishino, Y. *et al.* (2010). 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.
12. Xu, M. Q., Southworth, M. W., Mersha, F. B., Hornstra, L. J. & Perler, F. B. (1993). *In vitro* protein splicing of purified precursor and the identification of a branched intermediate. *Cell*, **75**, 1371–1377.
13. Paulus, H. (2000). Protein splicing and related forms of protein autoprocessing. *Annu. Rev. Biochem.* **69**, 447–496.
14. Wallace, C. J. A. (1993). The curious case of protein splicing: mechanistic insights suggested by protein semisynthesis. *Protein Sci.* **2**, 697–705.
15. Shao, Y., Xu, M. Q. & Paulus, H. (1996). Protein splicing: evidence for an N–O acyl rearrangement as the initial step in the splicing process. *Biochemistry*, **35**, 3810–3815.
16. Xu, M. Q., Comb, D. G., Paulus, H., Noren, C. J., Shao, Y. & Perler, F. B. (1994). Protein splicing: an analysis of the branched intermediate and its resolution by succinimide formation. *EMBO J*, **13**, 5517–5522.
17. Cooper, A. A. & Stevens, T. H. (1993). Protein splicing: excision of intervening sequences at the protein level. *BioEssays*, **15**, 667–674.
18. Shao, Y., Xu, M. Q. & Paulus, H. (1995). Protein splicing: characterization of the aminosuccinimide residue at the carboxyl terminus of the excised intervening sequence. *Biochemistry*, **34**, 10844–10850.
19. Van Roey, P., Pereira, B., Li, Z., Hiraga, K., Belfort, M. & Derbyshire, V. (2007). Crystallographic and mutational studies of *Mycobacterium tuberculosis recA* mini-inteins suggest a pivotal role for a highly conserved aspartate residue. *J. Mol. Biol.* **367**, 162–173.
20. Wood, D. W., Wu, W., Belfort, G., Derbyshire, V. & Belfort, M. (1999). A genetic system yields self-cleaving inteins for bioseparations. *Nat. Biotechnol.* **17**, 889–892.
21. Belfort, M. (1998). Inteins as antimicrobial targets: genetic screen for intein function. Patent 5795731.
22. Paulus, H. (2003). Inteins as targets for potential antimycobacterial drugs. *Front. Biosci.*, 8.
23. Derbyshire, V., Wood, D. W., Wu, W., Dansereau, J. T., Dalgaard, J. Z. & Belfort, M. (1997). Genetic definition of a protein-splicing domain: functional mini-inteins support structure predictions and a model for intein evolution. *Proc. Natl Acad. Sci. USA*, **94**, 11466–11471.
24. Wood, D. W., Derbyshire, V., Wu, W., Chartrain, M., Belfort, M. & Belfort, G. (2000). Optimized single-step affinity purification with a self-cleaving intein applied to human acidic fibroblast growth factor. *Biotechnol. Prog.* **16**, 1055–1063.
25. Hiraga, K., Derbyshire, V., Dansereau, J. T., Van Roey, P. & Belfort, M. (2005). Minimization and stabilization of the *Mycobacterium tuberculosis recA* intein. *J. Mol. Biol.* **354**, 916–926.

26. Perler, F. B., Olsen, G. J. & Adam, E. (1997). Compilation and analysis of intein sequences. *Nucleic Acids Res.* **25**, 1087–1093.

27. Pietrokovski, S. (1998). Modular organization of inteins and C-terminal autocatalytic domains. *Protein Sci.* **7**, 64–71.

28. Perler, F. B. (2002). InBase: the Intein Database. *Nucleic Acid Res.* **30**, 383–384 http://www.neb.com/neb/inteins.html; Updated July 21, 2010.

29. Ding, Y., Xu, M. Q., Ghosh, I., Chen, X., Ferrandon, S., Lesage, G. & Rao, Z. (2003). Crystal structure of a mini-intein reveals a conserved catalytic module involved in side chain cyclization of asparagine during protein splicing. *J. Biol. Chem.* **278**, 39133–39142.

30. Johnson, M. A., Southworth, M. W., Herrmann, T., Brace, L., Perler, F. B. & Wuthrich, K. (2007). NMR structure of a KlbA intein precursor from *Methanococcus jannaschii. Protein Sci.* **16**, 1316–1328.

31. Nichols, N. M., Benner, J. S., Martin, D. D. & Evans, T. C. J. (2003). Zinc ion effects on individual *Ssp* DnaE intein splicing steps: regulating pathway progression. *Biochemistry*, **42**, 5301–5311.

32. Senn, H. M. & Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angew. Chem., Int. Ed. Engl.* **48**, 1198–1229.

33. Warshel, A. (2003). Computer simulations of enzyme catalysis: methods, progress, and insights. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 425–443.

34. Lin, H. & Truhlar, D. G. (2007). QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **117**, 185–199.

35. Thompson, M. A. & Schenter, G. K. (1995). Excited states of the bacteriochlorophyll b dimer of *Rhodopseudomonas viridis*: a QM/MM study of the photosynthetic reaction center that includes MM polarization. *J. Phys. Chem.* **99**, 6374–6386.

36. Schoneboom, J. C., Lin, H., Reuter, N., Thiel, W., Cohen, S., Ogliaro, F. & Shaik, S. (2002). The elusive oxidant species of cytochrome P450 enzymes: characterization by combined quantum mechanical/molecular mechanical (QM/MM) calculations. *J. Am. Chem. Soc.* **124**, 8142–8151.

37. Gao, J. & Truhlar, D. G. (2002). Quantum mechanical methods for enzyme kinetics. *Annu. Rev. Phys. Chem.* **53**, 467–505.

38. Cui, Q., Elstner, M. & Karplus, M. (2002). A theoretical analysis of the proton and hydride transfer in liver alcohol dehydrogenase (LADH). *J. Phys. Chem. B*, **106**, 2721–2740.

39. Torrent, M., Vrenen, T., Musaev, D. G., Morokuma, K., Farkas, O. & Schlegel, H. B. (2002). Effects of the protein environment on the structure and energetics of active sites of metalloenzymes. ONIOM study of methane monoxygenase and ribonucleotide reductase. *J. Am. Chem. Soc.* **124**, 192–193.

40. Shemella, P. (2008). First principles study of intein reaction mechanisms. Diss. Rensselaer Polytechnic Institute.

41. Du, Z., Shemella, P. T., Liu, Y., McCallum, S. A., Pereira, B., Nayak, S. K. *et al.* (2009). Highly conserved histidine plays a dual catalytic role in protein splicing: a pK_a shift mechanism. *J. Am. Chem. Soc.* **131**, 11581–11589.

42. Voet, D., Voet, J. G. & Pratt, C. W. (2002). *Fundamentals of Biochemistry: Upgrade Edition.* John Wiley & Sons, Inc, New York, NY.

43. Garrett, R. H. & Grisham, D. (2009). *Biochemistry*, 4th edit Brooks/Cole, Boston, MA..

44. Leskovac, V. (2003). pp. 1–10, Kluwer Academic/Plenum Publishers, New York, Boston, Dordecht, London, Moscow.

45. Buskirk, A. R., Ong, Y. C., Gartner, Z. J. & Liu, D. R. (2004). Directed evolution of ligand dependence: small-molecule-activated protein splicing. *Proc. Natl Acad. Sci. USA*, **101**, 10505–105104.

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

47. Skretas, G., Meligova, A. K., Villalonga-Barber, C., Mitsiou, D. J., Alexis, M. N., Micha-Screttas, M. *et al.* (2007). Engineered chimeric enzymes as tools for drug discovery: generating reliable bacterial screens for the detection, discovery, and assessment of estrogen receptor modulators. *J. Am. Chem. Soc.* **129**, 8443–8457.

48. Frutos, E. A., Goger, M., Giovani, B., Cowburn, D. & Muir, T. W. (2010). Branched intermediate formation stimulates peptide bond cleavage in protein splicing. *Nat. Chem. Biol.* **6**, 527–533.

49. Shingledecker, K., Jiang, S. & Paulus, H. (2000). Reactivity of the cysteine residues in the protein splicing active center of the *Mycobacterium tuberculosis* RecA intein. *Arch. Biochem. Biophys.* **375**, 138–144.

50. Emsley, J., McEwan, P. A. & Gailani, D., Structure and function of factor XI. Blood 115, 2569—2577.

51. Kishi, T., Kato, M., Shimizu, T., Kato, K., Matsumoto, K., Yoshida, S. *et al.* (1999). Crystal structure of neuropsin, a hippocampal protease involved in kindling epileptogenesis. *J. Biol. Chem.* **274**, 4220–4224.

52. Debela, M., Hess, P., Magdolen, V., Schechter, N. M., Steiner, T., Huber, R. *et al.* (2007). Chymotryptic specificity determinants in the 1.0 Å structure of the zinc-inhibited human tissue kallikrein 7. *Proc. Natl Acad. Sci. USA*, **104**, 16086–16091.

53. Vijayalakshmi, J., Padmanabhan, K. P., Mann, K. G. & Tulinsky, A. (1994). The isomorphous structures of prethrombin2, hirugen-, and PPACK-thrombin: changes accompanying activation and exosite binding to thrombin. *Protein Sci.* **3**, 2254–2271.

54. Wu, W., Wood, D. W., Belfort, G., Derbyshire, V. & Belfort, M. (2002). Intein-mediated purification of cytotoxic endonuclease I-*Tev*I by insertional inactivation and pH-controllable splicing. *Nucleic Acids Res.* **30**, 4864–4871.

55. Davis, E. O., Sedgwick, S. G. & Colston, M. J. (1991). Novel structure of the RecA locus of *Mycobacterium tuberculosis* implies processing of the gene product. *J. Bacteriol.* **173**, 5653–5662.

56. Derbyshire, V., Kowalski, J. C., Dansereau, J. T., Hauer, C. R. & Belfort, M. (1997). Two-domain structure of the *td* intron-encoded endonuclease I-*Tev*I correlates with the two-domain configuration of the homing site. *J. Mol. Biol.* **265**, 494–506.

57. Maseras, F. & Morokuma, K. (1995). IMOMM—a new integrated ab initio plus molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *J. Comput. Chem.* **16**, 1170–1179.

58. Vreven, T., Morokuma, K., Farkas, O., Schlegel, H. B. & Frisch, M. J. (2003). Geometry optimization with QM/MM, ONIOM, and other combined methods. I.

Microiterations and constraints. *J. Comput. Chem.* **24**, 760–769.

59. Becke, A. D. (1993). Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652.

60. Hehre, W. J., Ditchfield, R. & Pople, J. A. (1972). Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **56**, 2257–2261.

61. Cornell, W. D., Cieplak, P., Bayley, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M. *et al.* (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197.

62. Olsson, M. H., Mavri, J. & Warshel, A. (2006). Transition state theory can be used in studies of enzyme catalysis: lessons from simulations of tunnelling and dynamical effects in lipoxygenase and other systems. *Philos. Trans. R. Soc. London, Ser. B*, **361**, 1417–1432.