Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 11084-11088, November 1994
Biochemistry

A proposed mechanism for the self-splicing of proteins

NEIL D. CLARKE

Department of Biophysics and Biophysical Chemistry, The Johns Hopkins University School of Medicine, Baltimore, MD 21205

Communicated by Paul Talalay, July 21, 1994

ABSTRACT **Intervening protein sequences, called inteins, are intronlike elements that are removed posttranslationally, apparently by self-splicing. The conserved and essential residues of precursor proteins consist of an asparagine as the last residue of the intein and a hydroxyl- or thiol-containing residue immediately following both splice junctions. Evidence for a branched intermediate has been reported [Xu, M.-Q., Southworth, M., Mersha, F., Hornstra, L. & Perler, F. (1993) *Cell* 75, 1371–1377]; however, the chemical nature of the branched structure is unclear. I propose a mechanism that includes the formation of a branched structure, provides an explanation for the reversal of branch formation observed at high pH, and accounts for each of the essential amino acids. The branched structure is formed by nucleophilic attack of the asparagine side chain on the N-terminal splice junction. The nature of this branched structure is a distinguishing feature of the model and can be experimentally tested.**

Proteins and nucleic acids can undergo rearrangements and modifications that add significant complexity to the decoding of genetic information. A striking example is the posttranslational splicing of precursor proteins to remove intronlike insertions of amino acids called inteins [the nomenclature used here follows the convention recently proposed by several workers in the field (1); see refs. 1–4 for reviews]. Inteins are not removed from RNA transcripts, but are instead translated in-frame as part of the protein in which they are inserted. The intein is subsequently removed in a self-splicing reaction.

Although only a handful of inteins have been identified, they have been discovered in eubacteria, archaea, and eukaryotes (5–7). They are roughly 50 kDa in size when spliced from the precursor and have sequence similarity to intron-encoded endonucleases found in group I introns. For several inteins, site-specific DNA endonuclease activity has been directly demonstrated; this activity is evidently related to the genetic mobility of the intein coding sequences (4, 8). There is no evidence, however, that endonuclease activity is related to the protein splicing reaction that removes the intein itself from the precursor protein.

The splice junctions of all inteins are closely related, and mutagenesis data indicate that three amino acids or amino acid types are required at precise positions at the junctions: the last residue of the intein must be asparagine, and a hydroxyl or thiol residue (serine, threonine, cysteine) must be at the C-terminal side of each of the two junctions (Fig. 1) (7, 10–12). The conserved hydroxyl or thiol residue at the N-terminal junction is thus the first residue of the intein; the hydroxyl or thiol residue at the C-terminal junction is the first residue of the protein fragment that follows the intein. Fragments of the precursor protein that become part of the mature protein are called exteins, and this particular extein is referred to as the C-extein because it is carboxyl-terminal to the intein (1).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

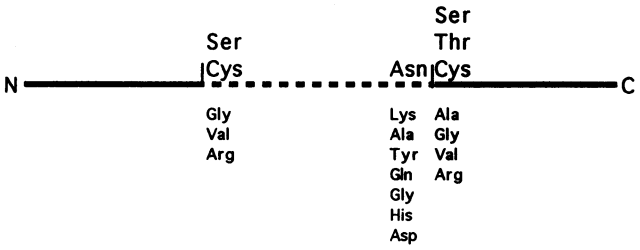

FIG. 1. Schematic of a precursor protein with the intein drawn as a dashed line. Amino acids listed above the line are those found at each position among the six known intein sequences. Asparagine is the last residue of the intein. Serine, threonine, and cysteine are the residues immediately C-terminal to each of the two splice junctions (see ref. 9 and the references therein for more complete sequence information). Amino acids listed below the line represent nonfunctional substitutions identified in mutagenesis experiments (7, 10–12). Excluded from this list are thiol or hydroxyl substitutions that have been found to be nonfunctional.

Histidine is conserved at the penultimate intein position, but it is not essential for splicing (10).

The evidence that inteins are removed posttranslationally rather than by mRNA splicing is compelling. While sequence changes that alter the amino acid sequence frequently result in a loss of splicing, mutations that alter only the nucleic acid sequence have no effect, even if those mutations overlap the splice junction (6, 10). The structural integrity of the intein polypeptide is apparently necessary for its own removal: substantial deletions within the intein decrease or abolish splicing, even if the correct reading frame is maintained (7). Frameshift mutations within the intein knock out splicing activity, but restoration of the reading frame by a second nearby mutation can restore it (7). In contrast, flanking sequences in the mature protein are unnecessary for splicing, with the exception of the hydroxyl or thiol residue that immediately follows the C-terminal splice junction (7, 9).

The recent report by Xu *et al.* (9) of *in vitro* splicing by a substantially purified chimeric precursor protein provides strong evidence that posttranslational splicing can occur by a self-splicing mechanism. The precursor protein, given the acronym MIP, is a fusion protein containing a maltose-binding domain (the N-extein), an intein from a DNA polymerase gene from *Pyrococcus*, and a domain of paromyosin (the C-extein). MIP was rapidly purified on a maltose-binding column and by FPLC (fast protein liquid chromatography) under conditions that slow the reaction (high pH and low temperature). A decrease in pH and an elevation of temperature resulted in formation of the free intein and mature protein (correctly spliced N-extein and C-extein) as the major products. The relative purity of the protein suggests that splicing may be an autocatalytic process.

The splicing of MIP *in vitro* can take several hours, permitting the detection of a putative intermediate. The intermediate was detected as an anomalously migrating band on SDS/polyacrylamide gels that appears and then disappears in the course of the reaction. The aberrant band contains all three fragments of the chimeric protein, but amino-terminal sequencing of the protein yielded roughly equal amounts of two

N-termini: that of the N-extein, which is also the N-terminus of the precursor and mature proteins, and that of the intein. Thus, it appears that there is a branched intermediate in which the peptide bond between the N-extein and the intein has been hydrolyzed but in which both the N-extein and the intein remain tethered to the C-extein. The chemical nature of the branch junction has not been elucidated. One clue to its identity is that high pH causes the apparent reversal of the branched structure to a precursor-sized molecule (9).

I propose here a mechanism for the self-splicing of proteins that explains the formation and reversibility of a branched structure, is chemically reasonable, and provides a rationalization for each of the essential amino acids. The branched structure predicted by this mechanism differs from that proposed by Xu *et al.* (9), and, in principle, can be readily distinguished experimentally.

Proposed Mechanism for Self-Splicing of Proteins

The self-splicing of a protein requires the hydrolysis of two peptide bonds and the formation of one new one. The mechanism I propose is outlined schematically in Fig. 2. In the interests of clarity, hypothetical proton donors and acceptors, including the conserved but nonessential histidine, are not shown. Except where noted, serine at the splice junctions is meant to represent serine, threonine, or cysteine. The mechanism, in outline, is as follows.

(*i*) The side chain of the invariant asparagine makes a nucleophilic attack on the peptide bond at the N-terminal splice junction, resulting in a branched structure with the experimentally observed N termini. The side-chain nucleophile could be either the amide nitrogen or the amide carbonyl. Attack by the amide nitrogen results in an imide with the carbonyl of the scissile peptide bond, while attack of the carbonyl results in an ester. Fig. 2 (structures I and II) shows the imide formation, but the reactions to form and break down the ester are similar to those shown for the imide. Fig. 3*A* shows the branched structures resulting from both imide and ester formation. This is the critical, distinguishing feature of this model. Subsequent steps are similar or identical to steps in previously proposed mechanisms, particularly those of Xu *et al.* (9), Cooper *et al.* (10), and Wallace (3).

(*ii*) The serine at the N-terminus of the intein attacks the branched structure imide (or ester) to form an ester between the serine side chain and the carbonyl derived from the hydrolyzed peptide bond. The result is a linear molecule in which the N-terminal splice junction has been activated for peptide bond formation by creation of an ester (Fig. 2, structures II and III). The asparagine side chain is free to play a role in step *iii*.

(*iii*) The peptide bond at the C-terminal splice junction is hydrolyzed, possibly by succinimide formation by the adjacent asparagine. Attack of the asparagine side chain on its own backbone carbonyl is shown very schematically in Fig. 2, structure III, and in greater detail in Fig. 4, reaction 2. The adjacent serine side chain can facilitate this reaction. Alternative hydrolysis reactions are possible.

(*iv*) The free amino group of the C-extein performs an aminolysis reaction on the ester at the N-terminal junction. The result is mature protein and free intein (Fig. 2, structures IV and V).

The Nature of the Branched Structure and Predicted Products of Hydrolysis

The distinguishing feature of this model is the nature of the branched structure. Fig. 3*A* shows the two alternative, but similar, branched structures proposed here, while Fig. 3*B* shows the rather different branched structure proposed by Xu *et al.* (9). The structures proposed here consist of the intein

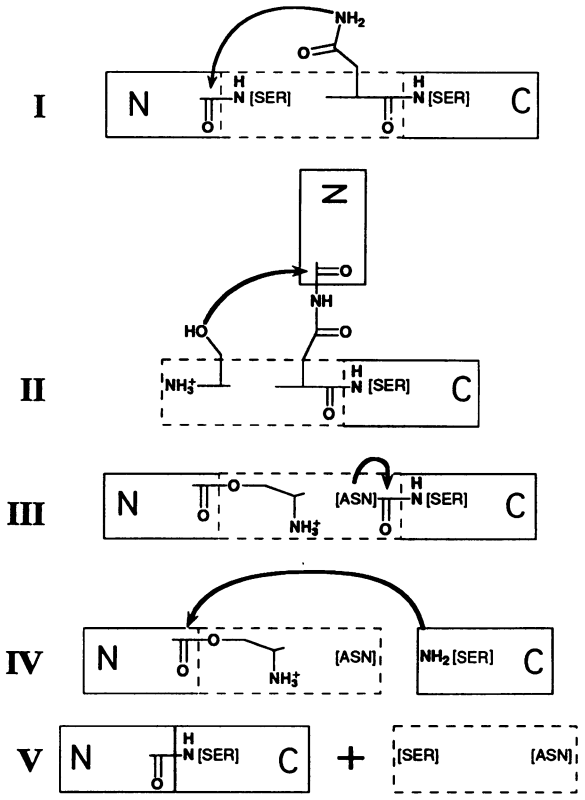

FIG. 2. Schematic of proposed mechanism. The N-terminal (N) and C-terminal (C) exteins are drawn as solid-lined boxes and the intein is drawn as a dash-lined box. Only the relevant amino acid side chains in each panel are shown. In cases where the side chain is not explicitly shown the position of required residues is indicated by the name of the residue. The use of serine in this figure is meant to represent any hydroxyl or thiol residue. All peptide, imide, and ester bonds linking the N-extein, C-extein, and intein are shown. Structure I, attack of asparagine on the N-terminal splice junction. Structure II, the branched structure that results from attack by asparagine. The branch linkage is then attacked by the N-terminal serine side chain. Structure III, the linear molecule that results from attack by serine on the asparagine branch link. The molecule has a serine-linked ester in place of a peptide bond at the N-terminal splice junction. This structure also indicates autocleavage at the C-terminal junction by asparagine. A more detailed view of this reaction is shown in Fig. 4. Structure IV, attack by the free amino group of the C-extein on the ester at the N-terminal junction. Structure V, the final products. Not shown are hypothetical proton donors and acceptors that presumably play a role in deprotonating the amine in structure IV and in promoting imide and ester formation. As discussed in the text, the carbonyl oxygen of the side-chain amide of asparagine could be the nucleophile in structure I instead of the amine, resulting in an ester-linked branch structure instead of the imide. The rest of the reaction would be essentially the same.

and the C-extein remaining as a single polypeptide, with the N-extein linked to the invariant and essential asparagine by an imide or an ester. In the alternative proposed by Xu *et al.* (9) (Fig. 3*B*), the branched structure consists of the mature protein as a single polypeptide with the intein linked through an ester to the splice junction serine of the mature protein. Since both imides and esters are more susceptible to hydrolysis than are peptide bonds, hydrolysis of the branched structure and analysis of the resulting polypeptides should permit the nature of the branched structures to be resolved. Hydrolysis of the structure in Fig. 3*B* would yield mature protein and free intein, while hydrolysis of either of the intermediates favored here (Fig. 3*A*) would yield free N-extein and an intact intein–C-extein polypeptide. Inspection of published figures from the work of Xu *et al.* (9) seems to indicate a relatively large amount of side product corresponding to an intein–C-extein polypep-

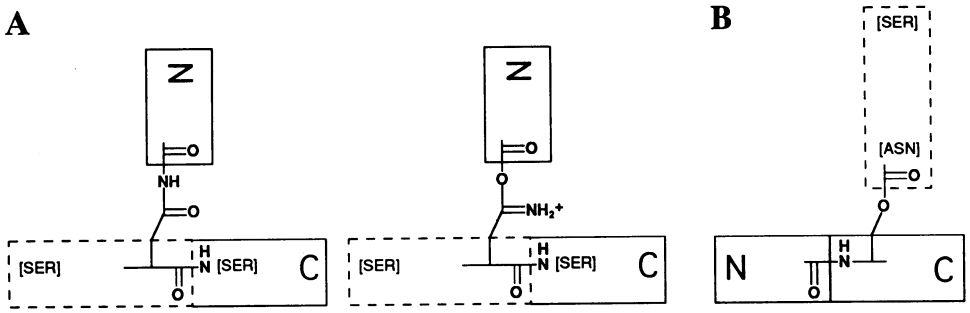

FIG. 3. (*A*) Alternative branched structure intermediates for the model proposed here. The imide linkage is on the left; the ester is on the right. (*B*) Branched structure intermediate of Xu *et al.* (9).

tide (9), consistent with this model. Such side products could arise by hydrolysis of some of the branched structure intermediate during preparation of samples for electrophoresis, although other explanations are equally plausible. Hydrolysis of purified branched structure intermediate will be required to resolve this question.

Reversibility of the Branched Structure and Its pH Dependence

The reversibility of the branched structure at pH 10 offers further support for the proposed mechanism. Reversal of branch formation in the model proposed here consists simply of an attack by the free amino group of the intein on the asparagine-linked imide or ester to re-form the peptide bond at the N-terminal splice junction (Fig. 5). Since branch formation is the first step in the proposed splicing mechanism, the reverse reaction regenerates the precursor protein. This process can be expected to be favored at high pH since (*i*) deprotonation of the N-terminal amine of the intein is necessary for the attack, (*ii*) the branch linkage itself is expected to be base-labile, and (*iii*) the precursor protein (ie., the product of branch reversal) is known to be stabilized by alkaline conditions (9, 10).

Another way to think about this is that formation of the branched structure formally results in net protonation of the protein at pH 7.0 (assuming normal pK_a values). This is because the peptide amide nitrogen at the splice junction becomes the protonated N terminus of the intein when the peptide bond is attacked by the asparagine; the peptide carbonyl remains linked to the asparagine. Reversal of the protonated branched structure, conversely, involves deprotonation of the protein. As a consequence, reversibility of the branched structure can be expected to be favored at high pH.

In contrast, similar considerations suggest that reversal of the branched structure of Xu *et al.* (9) would actually be unfavorable at high pH. In their mechanism, the immediate precursor to the branched structure is a linear molecule in which each of the splice-junction peptide bonds is replaced by an ester linkage between the side chain of the adjacent serine and the carbonyl oxygen of the hydrolyzed peptide bond. Each of the amide nitrogens from the hydrolyzed peptide bonds becomes a protonated primary amine, acting as a kind of side chain off this nonpeptide backbone (see the product of the reaction in Fig. 6 for a picture of one such esterified junction). The branched structure in their model is deprotonated relative to this precursor because one of the protonated amines is lost upon forming the new peptide bond at the N-extein–C-extein junction. Therefore, high pH in this case might be expected to favor the forward reaction from precursor to branched structure rather than the reverse reaction that was observed.

An important caveat to this argument is that the pK_a values of functional groups within the structure-specific environment of a protein can differ substantially from those in solution. Nevertheless, the available evidence appears most consistent with the branched structure proposed here.

It should be noted that structure III in the proposed mechanism (Fig. 2) is a full-length linear molecule that arises from the normal resolution of the branched intermediate. It is conceivable that the apparent reversibility of the branched structure is due to formation of this molecule rather than true reversal to the precursor protein. However, for reasons similar to those used above in support of the asparagine-linked structure, the formation and stability of structure III (Fig. 2) at high pH seem less likely than true reversal.

Imide Formation by Asparagine as the Basis for Protein Splicing

An important feature of the proposed mechanism is the role of the essential asparagine in forming the nonpeptide linkage

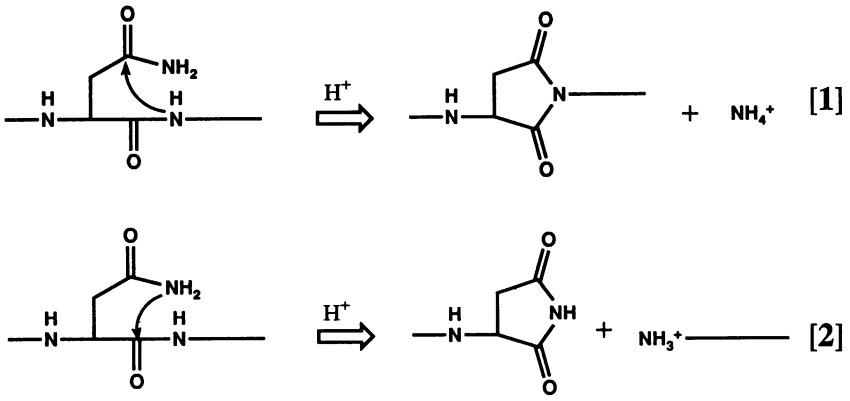

FIG. 4. Succinimide formation by asparagine. Reaction 1, succinimide formation leading to deamidation of asparagine. Reaction 2, succinimide formation resulting in polypeptide chain cleavage.

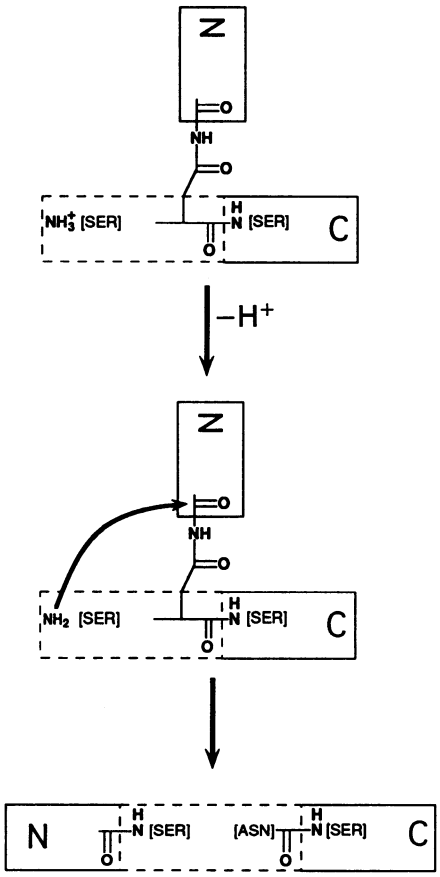

FIG. 5. Mechanism for branched structure reversibility. The deprotonation of the intein N-terminal amine at high pH is shown explicitly in the first step. Attack of the alkali-labile imide regenerates the precursor protein.

in the branched intermediate. This section and the next provide precedents to suggest that the proposed role is chemically reasonable.

The formation of imides between asparagine side chains and backbone atoms is well established (13–15). Two kinds of reactions can lead to imide formation. In the first, the backbone amide nitrogen attacks the carbonyl of the side-chain amide, resulting in a displacement of the side-chain nitrogen (Fig. 4, reaction 1). Formation of this succinimide is the first step in deamidation of asparagine to aspartic acid. The second way in which an imide can be formed between an asparagine and the backbone is by attack of the side-chain amide nitrogen on the carbonyl of the backbone amide (Fig. 4, reaction 2). In this case, it is the backbone amide nitrogen (and the rest of the polypeptide C-terminal to it) that is displaced. It is this type of imide formation that could give rise to the branched structure of Figs. 2 and 3*A*.

Although backbone cleavage as a result of asparagine–imide formation has only been observed at peptide bonds immediately C-terminal to the side chain, there is nothing in the chemistry itself that would suggest such a reaction cannot occur at peptide bonds remote in the sequence. In a normal protein the stereochemistry of the reaction is such that cleavage by asparagines at an adjacent peptide bond is clearly much more favorable than at any other site. It must be remembered, however, that intein sequences have evolved to perform the self-splicing reaction. As in any other protein-catalyzed reaction, the orientation of side chains in the protein is critical, and there is no obvious reason why an asparagine residue could not be positioned favorably for attack on the N-terminal splice junction.

An Asparagine Ester as an Alternative Branch Linkage

Nucleophilic attack by the invariant asparagine residue on the N-terminal splice junction could proceed through formation of either an imide or an ester. The imide has been favored here primarily because succinimides are quite alkali labile (consistent with the reversibility of the branched structure) and because of the ample evidence for succinimide formation in the deamidation of asparagines. However, esters can also be hydrolyzed under alkaline conditions, and there is some precedent for nucleophilic attack by the amide oxygen of the asparagine side chain. 3-Ketosteroid Δ^5-isomerase from *Pseudomonas testosteroni* forms a covalent bond with an acetylenic suicide inhibitor, and a bound active-site peptide has been isolated (16). Analysis of a rearranged peptide formed upon release of the inhibitor by mild acid hydrolysis suggested that the nature of the covalent cross-link is an imido ester formed by attack of the asparagine side-chain oxygen on a conjugated double bond of the steroid inhibitor (17). Interestingly, in light of these possible mechanistic similarities, the active site asparagine in this enzyme is followed by serine.

As shown in Fig. 3*A*, the branched structures formed during self-splicing by nucleophilic attack of asparagine would be similar whether the attacking nucleophile is the amide nitrogen or oxygen. Both completion of the splicing reaction and reversal of branch formation can proceed for the ester-linked branch in a manner similar to that shown for the imide-linked branch (Figs. 2 and 4).

Post-Branch-Formation Reactions

All of the steps in Fig. 2 following formation of the branched structure are hypothetical, as there is not yet any experimental evidence that directly addresses the question of what happens after branch formation. Fig. 2 shows one possible scenario for the post-branch-formation reaction. I will limit my discussion of these steps because there are similar or identical counterparts discussed in detail in other published mechanisms (3, 9, 10).

The reaction that breaks down the branched structure results in esterification of the N-intein to a serine side chain (Fig. 2, structure III). The ester is a reactive linkage that later can be attacked by the amino group of the C-extein to form the new peptide bond of the mature protein (Fig. 2, structure V). The same serine-linked ester is invoked in the mechanisms of Wallace (3) and Xu *et al.* (9), although in those mechanisms the ester is formed by direct attack of the serine side chain on the preceding peptide bond in a reaction called an *N–O*-acyl shift (Fig. 6).

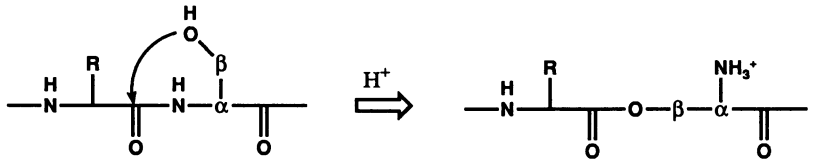

FIG. 6. *N–O*-acyl shift reaction. The R group of the amino acid N-terminal to the serine (e.g., asparagine) could facilitate this reaction. The resulting ester could be hydrolyzed to complete the breakage of the polypeptide chain.

The second reaction that must occur in some way is hydrolysis of the peptide at the C-terminal junction. The mechanism shown in Fig. 2 is succinimide formation by asparagine, which was discussed above in the context of asparagine reactivity. This reaction was proposed as a part of a protein splicing reaction by Cooper *et al.* (10). The adjacent serine may promote this reaction, as it is known that peptides containing asparagine-serine undergo autocleavage much more readily than asparagine-containing peptides that lack an adjacent serine (18). An alternative mechanism for C-terminal splice junction peptide cleavage is an *N–O*-acyl shift (3, 9).

The final step in the reaction (aminolysis of the ester at the N-terminal junction to form a new peptide bond) is shown schematically in Fig. 2 as a reaction distinct from the cleavage of the C-terminal junction, although it could occur in a concerted manner with previous steps. The depiction of the C-extein as a free species lacking covalent bonds to the rest of the protein is not meant to imply that it is free in solution. All fragments of the protein are likely to be tightly bound until the mature protein is formed.

Discussion

The discovery by Xu *et al.* (9) of a branched structure has provided the most important clue to date concerning the mechanism of protein splicing. The distinguishing feature of the model proposed here is the formation of a branched structure between the invariant asparagine and the N-extein. This proposed structure provides an explanation for the following observations:

(*i*) Formation of the branched structure appears to be reversible at high pH. The model discussed here provides a straightforward explanation for reversibility and is consistent with the observed pH dependence of this reaction.

(*ii*) The branched structure identified experimentally includes each of the three fragments of the precursor protein (N-extein, C-extein, and intein) and has free N-termini corresponding to both the intein and the N-extein. Both this model and that of Xu *et al.* (9) share these properties.

(*iii*) Asparagine at the C-terminus of the intein is conserved, and no other amino acid has been found to functionally substitute for it. Imide or ester formation at the N-terminal splice junction provides a very strong rationalization for this requirement.

A subtle feature of intein sequences not explained by this and other models is the fact that all known inteins have either a hydroxyl residue at both splice junctions or a cysteine at both junctions. Whether this is functionally significant or is in some way a vestige of the genetic transmission of the intein is unclear. Crystallographic studies of inteins should offer important insights into the self-splicing mechanism and the evolution of these sequences.

I thank numerous colleagues for helpful discussions during several presentations of this mechanism and Jeremy Berg for comments on an early version of this paper. I thank an anonymous reviewer for several corrections and helpful comments. This work was supported by the Markey Center for Macromolecular Structure and Function at The Johns Hopkins University.

1. Perler, F., Davis, E., Dean, G., Gimble, F., Jack, W., Neff, N., Noren, C., Thorner, J. & Belfort, M. (1994) *Nucleic Acids Res.* **22,** 1125–1127.
2. Shub, D. & Goodrich-Blair, H. (1992) *Cell* **71,** 183–186.
3. Wallace, C. (1993) *Protein Sci.* **2,** 697–705.
4. Doolittle, R. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5379–5381.
5. Kane, P., Yamashiro, C., Wolczyk, D., Neff, N., Goebl, M. & Stevens, T. (1990) *Science* **250,** 651–657.
6. Perler, F., Comb, D., Jack, W., Moran, L., Qiang, B., Kucera, R., Benner, J., Slatko, B., Nwankwo, D., Hempstead, S., Carlow, C. & Jannash, H. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 5577–5581.
7. Davis, E., Jenner, P., Brooks, P., Colston, M. & Sedgwick, S. (1992) *Cell* **71,** 201–210.
8. Gimble, F. & Thorner, J. (1992) *Nature (London)* **357,** 301–306.
9. Xu, M.-Q., Southworth, M., Mersha, F., Hornstra, L. & Perler, F. (1993) *Cell* **75,** 1371–1377.
10. Cooper, A., Chen, Y.-J., Lindorfer, M. & Stevens, T. (1993) *EMBO J.* **12,** 2575–2583.
11. Hodges, R., Perler, F., Noren, C. & Jack, W. (1992) *Nucleic Acids Res.* **20,** 6153–6157.
12. Hirata, R. & Anraku, Y. (1992) *Biochem. Biophys. Res. Commun.* **188,** 40–47.
13. Geiger, T. & Clarke, S. (1987) *J. Biol. Chem.* **262,** 785–794.
14. Stephenson, R. & Clarke, S. (1989) *J. Biol. Chem.* **264,** 6164–6170.
15. Wright, H. (1991) *Crit. Rev. Biochem. Mol. Biol.* **26,** 1–52.
16. Penning, T. & Talalay, P. (1981) *J. Biol. Chem.* **256,** 6851–6858.
17. Penning, T., Heller, D., Balasubramanian, T., Fenselau, C. & Talalay, P. (1982) *J. Biol. Chem.* **257,** 12589–12593.
18. Tyler-Cross, R. & Schirch, V. (1991) *J. Biol. Chem.* **266,** 22549–22556.