Review

Protein splicing – the lengths some proteins will go to

Elaine O. Davis & Peter J. Jenner

Laboratory for Leprosy and Mycobacterial Research, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, England

Received 18 February 1994; accepted in revised form 18 February 1994

Key words: protein splicing, post-translational modification, intein, extein, review

Abstract

We review the recently discovered phenomenon of protein splicing which is the excision of an internal protein sequence at the protein level rather than at the RNA level. The means by which examples of protein splicing have been identified are described, and the similarities of the internally spliced protein products (or inteins) are discussed. Comparisons are made between inteins and group I RNA introns. We describe the evidence supporting excision of inteins by a post-translational autocatalytic reaction of a full length polypeptide precursor, rather than by RNA splicing. An examination is made of some of the proposed mechanism schemes and the supporting them presented.

Introduction

The recent discovery of protein splicing has extended the range of methods by which the flow of information from DNA to mature protein can be modified. Far fewer examples of protein splicing than RNA splicing are known, but all share some common characteristics. Protein splicing can be defined as the precise removal of an embedded in-frame internal protein sequence from a polypeptide precursor, with the ligation of the two remaining external protein sequences to produce an active mature protein. Recently it has been proposed that the internal sequence be called the intein, and the external protein sequences be known as the exteins (Fig. 1); terms that are analogous to RNA splicing components introns and exons (Perler et al. in press).

Instances of protein splicing

Protein splicing was first identified in the VMA1/TFP1 allele of Saccharomyces cerevisiae which encodes the 69 kDa subunit A of the yeast vacuolar ATPase. Hirata et al. (1990) and Kane et al. (1990) observed that the VMA1/TFP1 gene contains an open reading frame sufficient to encode a 1071 amino acid, 119 kDa protein, much larger than the size of other known ATPase subunits. Sequence analysis revealed that a high degree of homology with these ATPases resided in the N- and C-terminal regions alone, while the intervening spacer domain possessed very little homology with ATPases. Moreover, Northern blot analysis showed that the gene transcript was 3.7 kb, sufficient to code for the entire 119 kDa polypeptide. Other experiments using Western blotting and mutagenesis led to the conclusion that the intervening protein was removed after translation. Both groups of workers came to the conclusion that the intein was excised by a rapid autocatalytic reaction although the evidence for this was largely circumstantial.

Subsequently, protein splicing has been identified in the DNA polymerase of the archaeabacteria Thermococcus litoralis and Pyrococcus species strain GB-D (Perler et al. 1992; F. Perler, personal communication) and RecA from Mycobacterium tuberculosis and M. leprae (Davis et al. 1991; Davis et al. 1992; Davis et al. 1994). Recently an intein has also been found in the vacuolar ATPase subunit from Candida tropicalis at the same position as that in S. cerevisiae and in this case an RNA intron is present in addition in the same gene (Gu et al. 1993). T. litoralis differs from the other...
Translation

Polypeptide precursor

Splicing

Extein

Intein

Fig. 1. Diagrammatic representation of protein splicing. The complete RNA transcript is translated into a polypeptide precursor which contains an in-frame internal protein sequence (intein), shown in light, embedded between the N- and C-terminal external protein sequences (exteins), shown in dark. In the splicing reaction the intein is excised, and the two exteins are ligated to form an active mature protein.

examples in having two inteins, originally designated IVPS1 and IVPS2, which in the native organism are spliced out to leave the mature polymerase, known as Vent DNA polymerase (Perler et al. 1992; Hodges et al. 1992). The intact gene was not stable when cloned in E. coli unless intein-1, the amino-terminal intein, was first deleted, when intein-2 was spliced out to give active polymerase. In a construct containing intein-1 but truncated a short distance into intein-2, no splicing of intein-1 was found in E. coli. The Pyrococcus polymerase contains just one intein which is at the same location as intein-1 of Thermococcus and shows homology to it, but in this case splicing occurs in E. coli. In contrast to the examples in the two yeasts and the two archaeabacteria, the two mycobacterial inteins are located in different positions within the RecA protein. Although the M. tuberculosis intein is spliced very efficiently in E. coli, the M. leprae intein is not spliced in this heterologous system but only in M. leprae itself. Despite these minor differences and in contrast to the diversity of the organisms and types of proteins in which protein splicing occurs, the inteins of these examples share a number of similar properties, and these will be discussed in more detail below.

Identification of inteins is complicated by the lack of general homology between them

Identifying examples of inteins is reliant on sequence analysis, either by finding an internal breakdown in homology with other equivalent genes followed by a downstream resumption of the homology, or by inference from an anomalous molecular weight from PAGE data compared with that deduced from the open reading frame. It is perhaps because of this that relatively few examples have been found so far. In all the cases identified the disruption of a highly conserved protein sequence allowed the presence of an intein to be deduced. A good example is that of M. tuberculosis RecA (Davis et al. 1991; Davis et al. 1992). More than 20 recA genes have been cloned from gram-negative bacteria, and all possess a striking degree of similarity. After cloning and sequencing the gene from M. tuberculosis, comparison with other RecA’s showed a remarkable discrepancy both in the size of the open reading frame and in an interruption of the high degree of homology with the E. coli RecA sequence (Fig. 2) which resided solely in the N- and C-terminal segments. The protein encoded by the gene’s open reading frame would have a predicted molecular mass of 85 kDa, but the apparent molecular size as seen by Western blotting was 38 kDa, similar to other RecA’s. The amino acid sequence at the proposed splice site junctions showed a considerable degree of conservation with those of the VMA1 protein (Fig. 2) (Davis et al. 1991; Hirata et al. 1990) being especially significant at the second splice site, although the remaining sequence had very little homology.

Conservation at the splice sites is found to varying degrees in all the known examples of inteins (Fig. 3) and suggests a common mechanism of excision of the intein. What is noticeable though, is that while the extein sequences are highly conserved with the sequences of their other known homologues, the inteins possess much less homology with each other, with the single exception of Tli-intein-1 and the Pyrococcus intein. The overall amino acid identity amongst the sev-