© 1994 Oxford University Press
Nucleic Acids Research, 1994, Vol. 22, No. 7 1125–1127

Protein splicing elements: inteins and exteins — a definition of terms and recommended nomenclature

Francine B.Perler*, Elaine O.Davis[1], Gary E.Dean[2], Frederick S.Gimble[3], William E.Jack, Norma Neff[4], Christopher J.Noren, Jeremy Thorner[5] and Marlene Belfort[6]
New England Biolabs, Inc., Beverly, MA 01915, USA, [1]National Institute for Medical Research, Mill Hill, London, UK, [2]University of Cincinnati College of Medicine, Cincinnati, OH 45267-0524, [3]Institute of Biosciences and Technology, Texas A & M University, Houston, TX 77030-3303, [4]Sloan-Kettering, Molecular Biology Program, New York, NY 10021, [5]University of California, Berkeley, CA 94720 and [6]Wadsworth Center for Laboratories and Research, NY State Department of Health, Albany, NY 12201-0509, USA

Received November 24, 1993; Revised and Accepted February 28, 1994

INTRODUCTION

Several archaeal, eubacterial and eucaryotic genes have been identified with in-frame insertions that are excised at the protein level, not at the RNA level (1–13, 24, reviewed in 14–21). This process is termed **protein splicing**. Initially, a single precursor polypeptide is synthesized. The intervening protein sequence is then excised from within the precursor, and the flanking protein sequences are joined. Thus, protein splicing results in the production of two proteins from a single primary translation product, the internal protein and the protein formed by the joining of the external sequences. The removal of the internal segment, concomitant with the formation of a normal peptide bond joining the external polypeptide sequences, distinguishes protein splicing from simple autoproteolysis (2). The rapid production of the mature products suggests that protein splicing is very efficient. The protein precursor rarely accumulates, even when the native gene is expressed in heterologous systems, both *in vivo* and *in vitro* (1–12). Evidence to date suggests that protein splicing is autocatalytic.

DEFINITIONS

A uniform nomenclature for the elements involved in this process is needed. The nomenclature suggested here is patterned after that of introns and RNA splicing (21,22). Accordingly, we propose calling the process **protein splicing** and the primary translation product, the **precursor protein** (Figure 1). We suggest calling the ligated product the **mature protein**, the **ligated protein**, or the **spliced protein**. We suggest calling the internal protein sequence, the **intein**, derived from *in*ternal pro*tein* sequence. We propose that the *ex*ternal pro*tein* sequences in the precursor be called **exteins**.

An **intein** is defined as a protein sequence embedded *in-frame* within a precursor protein sequence that is spliced out during maturation. Intein refers to the intervening sequence both when it is part of the precursor and after it has been liberated by splicing (Figure 1). The spliced intein should be referred to as the **free intein** to distinguish it from the **fused intein** present in the precursor. An intein is analogous to an RNA intron. This nomenclature should replace that used in previous publications, where inteins have been referred to as 'spacers', 'protein introns', 'protein inserts' or 'intervening protein sequences (IVS and IVPS)'.

The protein sequences that flank the intein, and that are ligated to form the mature product, are defined as the **exteins**. Exteins are analogous to RNA exons. This nomenclature should replace that used in previous publications, where exteins were called 'external protein sequences' or 'EPSs'.

We further suggest that the full intein name should include both a genus and species designation (which can be abbreviated with the standard 3 letter genus/species convention) and a host gene designation. For example, the *S.cerevisiae VMA* intein would be called, '*Sce VMA* intein' (Table 1). If a protein contains more than one intein, they should be numbered sequentially beginning at the N-terminus of the precursor (Figure 1). Thus, the first intein should be designated **intein-1** and the second, **intein-2**, etc. Exteins can be numbered **extein-1, extein-2, extein-3,** etc., beginning at the N-terminus of the precursor. Alternatively, the extein on the N-terminal side of any intein can be referred to as the **upstream extein**, **N-terminal extein** or **N-extein** and the extein on the C-terminal side of the intein can be referred to as the **downstream extein**, **C-terminal extein**, or **C-extein**.

Finally, we suggest that the nucleotide sequence which encodes the intein portion of the precursor gene or RNA be called the **intein coding sequence**. The name of the mature product should be used for the entire gene encoding inteins and exteins. For example, the *M.tuberculosis recA* gene encodes the RecA precursor protein, the mature RecA protein and the *Mtu recA* intein.

*To whom correspondence should be addressed

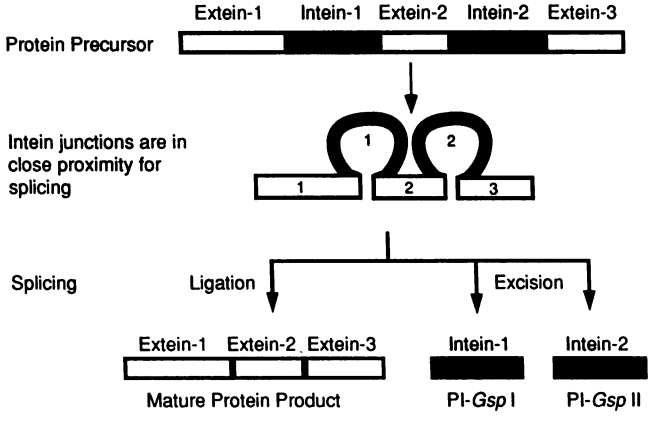

Figure 1. Protein splicing results in the production of two or more proteins from a single precursor protein. Protein splicing requires the precise excision of the intein(s) and the ligation of the exteins to form the mature product and the free intein(s). In the example depicted, two inteins are present in a single precursor protein. If multiple inteins are present in a single precursor, they should be numbered sequentially beginning at the N-terminus. If endonuclease activity has been demonstrated, the inteins should also be named according to protein splicing endonuclease naming conventions (prefix, PI-, *G* representing the genus, *sp* representing the species, and I and II corresponding to the order of discovery of endonuclease activity).

Table 1. Properties of characterized inteins

New Name	Old Name(s)	Endonuclease	References
Sce VMA intein	Spacer/VDE	PI-*Sce*I	5,6,8,11
Mtu recA intein	Spacer		3, 4
Tli pol intein-1	*Tli* IVPS1	PI-*Tli*II	2, 10
Tli pol intein-2	*Tli* IVPS2	PI-*Tli*I	2, 10
Ctr VMA intein	Spacer		7
Psp pol intein	*Psp* IVPS1	PI-*Psp*I	13
Mle recA intein	Protein intron		24

Intein names should include an organism (3 letter genus/species abbreviation) and gene designation. If multiple inteins are present in a gene, inteins should be numbered in order of occurrence in the precursor protein, beginning at the N-terminus.

CHARACTERIZED INTEINS

A total of seven inteins have been identified in the following proteins: (a) the 69 kDa subunit of the vacuolar ATPase from *Saccharomyces cerevisiae* (1,2,5,6,8,9,11) and *Candida tropicalis* (7), (b) RecA from *Mycobacterium tuberculosis* (3,4) and *Mycobacterium leprae* (24 and D. Smith, Collaborative Research, personal communication), and (c) DNA polymerases from the archaeon *Thermococcus litoralis* (2 inteins, 10,12) and *Pyrococcus species* strain GB-D (1 intein, 13 and Perler, *et al.*, in preparation). In each case, an in-frame intein must be precisely removed from the protein precursor and the exteins joined to produce the active, mature protein. Table 1 lists the proposed names for each of these inteins.

M.leprae and *M.tuberculosis* RecA precursor proteins each have a single, unrelated intein present at different positions that are 46 amino acids apart in the uninterrupted RecA protein. *T.litoralis* DNA polymerase precursor has two unrelated inteins separated by 49 amino acids. The *Psp pol* intein is similar to the *Tli pol* intein-1 and is located at the same position within the DNA polymerase precursor. The two yeast ATPase precursors also have similar inteins located at the same position within the precursor. In many of the above cases, alleles without inteins have been found in isolates of the same or similar species, suggesting the possibility of lateral transmission (2,4,5,19–20,24 and Perler *et al.*, in preparation).

INTEINS AND HOMING ENDONUCLEASES

All of the known inteins have sequence similarity to homing endonucleases (reviewed in 14,15,17–20,22–23), but not all of them have been tested for endonuclease activity. Once endonuclease activity has been established, intein-derived endonucleases should be named according to the standard homing endonuclease nomenclature conventions, but preceded by the prefix '**PI-**' (**P**rotein **I**nsert (19) or **P**rotein **I**ntervening sequence) instead of the intron homing endonuclease prefix, 'I-' (22). Endonucleases are named by using the first letter of the genus followed by a 2 letter species abbreviation. Intein-derived endonucleases should be named as above and numbered in order of discovery. However, intein homing endonucleases should be numbered independently from RNA-derived homing endonuclease. To avoid confusion in naming successive endonucleases from the same organism, we suggest that new intein homing endonucleases be submitted to Dr Francine Perler who will keep a registry of intein homing endonuclease names.

Thus, the product of the *Psp pol* intein is called PI-*Psp*I. Similarly, the product of the *Sce VMA* intein (formerly called 'VDE' for VMA1-derived endonuclease) is now called PI-*Sce*I (19,20). The prefix 'PI-' is used to distinguish homing endonucleases derived from protein splicing elements from homing endonucleases derived from introns, and does not imply any differences in endonuclease function or properties. Otherwise, terms used for intron-derived homing endonucleases should also be used for homing endonucleases generated by protein splicing (20,22). Inteins should not be given endonuclease names until activity has been demonstrated since inteins that lack endonuclease activity may be discovered. The endonuclease name, such as PI-*Psp*I, may be used to refer to the free endonuclease or to the endonuclease sequence embedded in the precursor protein.

Endonucleases PI-*Tli*II and PI-*Psp*I, encoded by the intein alleles *Tli pol* intein-1 and *Psp pol* intein, respectively, are isoschizomers, cleaving the same DNA sequences (Perler *et al.*, in preparation). In the case of *T.litoralis* DNA polymerase, which is the only precursor with two inteins identified to date, endonuclease activity was first demonstrated for *Tli pol* intein-2 and only much later for *Tli pol* intein-1, resulting in *Tli pol* intein-1 encoding PI-*Tli*II and *Tli pol* intein-2 encoding PI-*Tli*I. This type of inconsistent numbering is inevitable, given that the order of inteins in a gene need not necessarily parallel the order of discovery of their endonuclease activity in that gene, let alone in that species.

INTEIN MOBILITY

The homing endonucleases found in many self-splicing introns have been shown to mediate intron mobility (19–22). Likewise, the *S.cerevisiae* intein has also been shown to be mobile (5). Therefore, as in the case of self-splicing introns, inheritance of inteins can be either horizontal (i.e., transmission by gene mobilization and insertion) or vertical (i.e., normal chromosomal transmission). Intein mobility, however, suggests that inteins found in the same location in different organisms are likely to be isoschizomers, as in the case of PI-*Psp*I and PI-*Tli*II.

In accordance with the homing intron nomenclature (22), the **homing site** is comprised of the sequences surrounding the point in the gene into which the intein coding sequence from a mobile intein is inserted. The **intein insertion site** is the junction of the extein coding sequences in alleles lacking inteins. **Intein homing** is the acquisition of an intein coding sequence at a specific site in a gene.

PROTEIN SPLICING IN FOREIGN CONTEXTS

The intein plus the first residue of the downstream extein are sufficient for protein splicing when cloned into foreign or target proteins (2, 4, 13). However, splicing of inteins in foreign contexts is sometimes less efficient than in the native protein, resulting in cleavage at single splice junctions or the excision of an intein in the absence of ligation of the exteins. **Excision** refers to the production of free intein from a precursor and can be independent of the production of the mature protein, whereas **protein splicing** refers to the excision of the intein coupled with ligation of the exteins. **Cleavage** refers to the breakage of the peptide bond at a single intein/extein junction. Therefore, in analyzing splicing data, it is important not to rely only on excision as a measure of protein splicing.

As more genes are sequenced, it is likely that more inteins will be discovered. We hope that this suggested nomenclature will provide uniformity to the field and will prove useful as more of these elements are found.

ACKNOWLEDGEMENTS

We thank Drs Yasujiro Anraku, Heidi Goodrich-Blair, Alan Lambowitz, Maurice Southworth, David Shub, Douglas Smith, Tom Stevens, and Ming-Qun Xu for helpful discussions.

REFERENCES

1. Bremer, M.C.D., Gimble, F.S., Thorner, J. and Smith, C.L. (1992) *Nucleic Acids Res.*, **20**, 5484.
2. Cooper, A.A., Chen, Y., Lindorfer, M.A. and Stevens, T.H. (1993) *EMBO J.*, **12**, 2575–2583.
3. Davis, E.O., Sedgwick, S.G. and Colston, M.J. (1991) *J. Bacteriol.*, **173**, 5653–5662.
4. Davis, E.O., Jenner, P.J., Brooks, P.C., Colston, M.J. and Sedgwick, S.G. (1992) *Cell*, **71**, 201–210.
5. Gimble, F.S. and Thorner, J. (1992) *Nature*, **357**, 301–306.
6. Gimble, F.S. and Thorner, J. (1993) *J. Biol. Chem.*, *268*, 21844–21853 .
7. Gu, H.H., Xu, J., Gallagher, M. and Dean, G.E. (1993) *J. Biol. Chem.*, **268**, 7372–7381.
8. Hirata, R., Ohsumi, Y., Nakano, A., Kawasaki, H., Suzuki, K. and Anraku, Y. (1990) *J. Biol. Chem.*, **265**, 6726–6733.
9. Hirata, R. and Anraku, Y. (1992) *Biochem. Biophys. Res. Comm.*, **188**, 40–47.
10. Hodges, R.A., Perler, F.B., Noren, C.J. and Jack, W.E. (1992) *Nucleic Acids Res.*, **20**, 6153–6157.
11. Kane, P.M., Yamashiro, C.T., Wolczyk, D.F., Neff, N., Goebl, M. and Stevens, T.H. (1990) *Science*, **250**, 651–657.
12. Perler, F.B., Comb, D.G., Jack, W.E., Moran, L.S., Qiang, B., Kucera, R.B., Benner, J., Slatko, B.E., Nwankwo, D.O., Hempstead, S.K., Carlow, C.K.S. and Jannasch, H. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 5577–5581.
13. Xu, M., Southworth, M.W., Mersha, F.B., Hornstra, L.J. and Perler, F.B. (1993) *Cell*, **75**,1371–1377.
14. Doolittle, W.F. and Stoltzfus, A. (1993) *Nature*, **361**, 403.
15. Doolittle, R.F. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 5379–5381.
16. Hendrix, R.W. (1991) *Current Biology*, **1**, 71–73.
17. Shub, D.A. and Goodrich-Blair, H. (1992) *Cell*, **71**, 183–186.
18. Wallace, C.J.A. (1993) *Protein Science*, **2**, 697–705.
19. Lambowitz, A.M. and Belfort, M. (1993) *Annu. Rev. Biochem.*, **62**, 587–622.
20. Mueller, J.E., Bryk, M., Loizos, N. and Belfort, M. (1994), Nucleases, Cold Spring Harbor Press, Cold Spring Harbor. pp. 111–143.
21. Krainer, A.R. and Maniatis, T. (1988), Transcription and Splicing, IRL Press, Oxfod. pp. 131–206.
22. Dujon, B., Belfort, M., Butow, R.A., Jacq, C., Lemieux, C., Perlman, P.S. and Vogt, V.M. (1989) *Gene*, **82**, 115–118.
23. Heitman, J. (1993), Genetic Engineering, Plenum Press, New York. pp. 57–108.
24. Davis, E.O., Thangaraj, H.S., Brooks, P.C., and Colston, M.J. (1994) *EMBO.J.* **13**,699–703