Available online at www.sciencedirect.com
ScienceDirect
Fungal Genetics and Biology 44 (2007) 153–179
FUNGAL GENETICS AND BIOLOGY
www.elsevier.com/locate/yfgbi

Review

The nuclear-encoded inteins of fungi

Russell T.M. Poulter, Timothy J.D. Goodwin, Margaret I. Butler *

Department of Biochemistry, University of Otago, P.O. Box 56, Dunedin, New Zealand

Received 25 July 2006; accepted 28 July 2006
Available online 13 October 2006

Abstract

An intein is a protein sequence embedded within a precursor protein that is excised during protein maturation. Inteins were first found encoded in the *VMA* gene of *Saccharomyces cerevisiae*. Subsequently, they have been found in diverse organisms (eukaryotes, archaea, eubacteria and viruses). The VMA intein has been found in various saccharomycete yeasts but not in other fungi. Different inteins have now been found widely in the fungi (ascomycetes, basidiomycetes, zygomycetes and chytrids) and in diverse proteins. A protein distantly related to inteins, but closely related to metazoan hedgehog proteins, has been described from Glomeromycota. Many of the newly described inteins contain homing endonucleases and some of these are apparently active. The enlarged fungal intein data set permits insight into the evolution of inteins, including the role of horizontal transfer in their persistence. The diverse fungal inteins provide a resource for biotechnology using their protein splicing or homing endonuclease capabilities.

Index descriptors: Intein; Protein splicing; Homing endonuclease; Hedgehog protein; Vacuolar ATPase (VMA); Pre-mRNA processing (PRP8); RNA polymerase; Horizontal gene transfer; Meiotic drive; Antibiotic targets

1. Introduction

An intein is a protein sequence embedded within a precursor protein sequence and excised during protein maturation, that is, post-translationally (Perler et al., 1994). Sequences encoding inteins occur sporadically within the protein coding sequences of eubacteria, archaea and eukaryotes. The intein coding sequence is translated, together with the surrounding sequence, to produce a precursor protein. The intein (internal protein) is then excised and the flanking protein sequences (exteins) are joined (spliced) by a normal peptide bond to form the functional protein. The excision of the intein and the subsequent ligation of the flanking exteins are catalysed by the intein itself (Xu et al., 1993; Chong et al., 1996). As a result of this 'protein splicing', two proteins (the spliced exteins and the free intein) are formed from a single precursor. In the strict sense the term intein refers to a protein molecule, but the gene segment encoding the intein is also often referred to as an intein. When the sequences of homologous genes are aligned an intein encoding sequence can be recognised as an insertion apparent in one or more of the homologues.

As well as catalysing their own protein splicing, many inteins also include site-specific homing endonucleases (Chevalier and Stoddard, 2001). These homing endonucleases belong to several distinct families, but the LAGLIDADG type is by far the most common in inteins. Homing endonucleases can initiate a gene conversion event that converts a cell heterozygous for the intein into a homozygote. The homing endonuclease recognises a long DNA target sequence corresponding to an unoccupied allelic site and generates a double strand break within the unoccupied target sequence. This break is repaired by the host repair machinery using as a template the allele containing the intein encoding sequence. This gene conversion results in the replication of the intein coding sequence into a specific site in a previously unoccupied allele. The occupied allele is no longer a target for the homing endonuclease because the target site is split by the insertion (Gimble and Thorner, 1992). This replication process, in which the intein encoding sequence behaves as a

* Corresponding author. Fax: +64 3 479 7866.
E-mail address: margi.butler@stonebow.otago.ac.nz (M.I. Butler).

1087-1845/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.fgb.2006.07.012

mobile element, will increase the frequency of the intein encoding sequence within the gene pool of the species. Unlike most mobile elements, a particular intein encoding sequence will occur just once in a haploid genome and at a specific genetic position ('home') in a defined gene.

A significant proportion of known inteins lack an endonuclease domain. These minimal protein-splicing elements (mini-inteins) are 130–200 amino acids in length with conserved sequence blocks at each end (Perler, 2000). There are several cases where a group of 'allelic' inteins (inteins found at the same insertion points in homologous proteins) comprises both full-length (homing endonuclease containing) and mini-inteins; for example, the intein in the ribonucleoside-diphosphate reductase, α subunit of *Pyrococcus furiosus* (Pfu RIR1-2) contains endonuclease motifs but *Methanothermobacter thermautotrophicus* contains a mini-intein (Mth RIR1) at this site.

Allelic inteins are usually highly similar in sequence and cluster together on phylogenetic trees (Gogarten et al., 2002; Butler et al., 2005). Non-allelic inteins usually have very little sequence similarity and their sequences are difficult to align. Some proteins contain more than one intein, however, these inteins are no more closely related to each other than to any other intein (Perler et al., 1997).

Both full-length inteins and mini-inteins almost invariably have four conserved splice-junction residues; a serine or cysteine at their N-termini, the dipeptide histidine–asparagine or histidine–glutamine at their C-termini, and a serine, threonine or cysteine following the downstream splice site (Fig. 1). Another highly conserved intein feature is the histidine in motif B/N3 (Fig. 1). As further inteins have been identified, however, an ever wider range of sequence variation within them has been recognised (Perler, 2005), leaving the first residue of the C-extein as the only sequence feature common among non-allelic inteins (Dassa and Pietrokovski, 2005).

The biochemical details of the protein splicing mechanism in canonical inteins (i.e., those which contain the conserved features shown in Fig. 1) were described by a number of research groups during the early 1990s (reviewed in Perler et al., 1997; Paulus, 2000; Mills and Paulus, 2005). To briefly summarise, protein splicing proceeds via four steps, described here for an intein with a cysteine residue following each splice junction. The initial step is the attack by the nucleophilic side chain of the cysteine at the N-terminus of the intein on the upstream peptide bond to form an ester intermediate. The N-extein is then transferred to the side chain of the cysteine immediately downstream of the intein (the C-extein cysteine). This branched structure is resolved by cyclisation of the intein C-terminal asparagine to a succinimide ring and cleavage of the C-terminal splice junction. At this stage, the exteins are linked by an ester bond. The final step is the acyl rearrangement that links the exteins with a peptide bond (Perler, 1998; Evans and Xu, 1999).

The splicing domains of inteins are horseshoe shaped with twofold symmetry (Moure and Quiocho, 2005). This domain has been shown to have the same fold as that found in hedgehog signalling domains (Hall et al., 1997) indicating that these two protein types share a common ancestor; the fold is now termed the Hint (hedgehog intein) domain

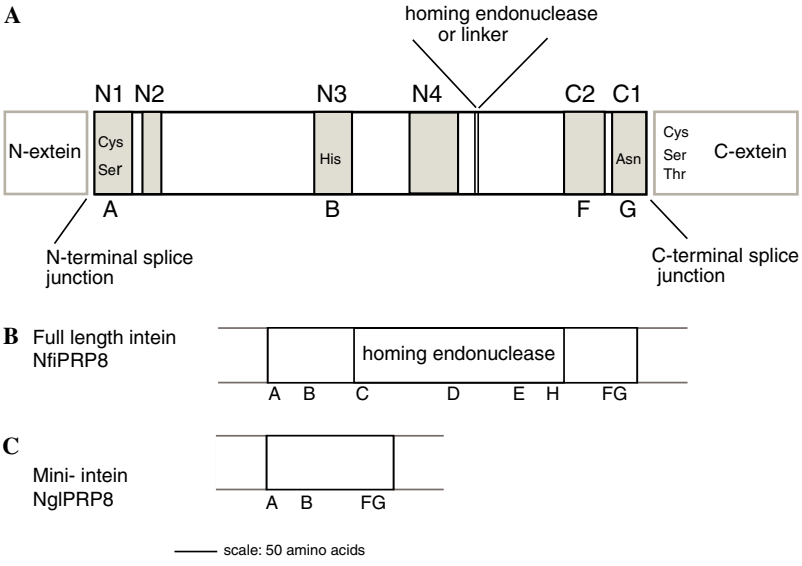

Fig. 1. Conserved elements in a canonical intein. (A) Conserved elements in a canonical intein. The shaded regions represent conserved motifs; they are identified by two different nomenclatures: above the intein box are the terms (N1, N2, N3, N4, C2, C1) suggested by Pietrokovski (1998a), below the intein box are terms (A, B, F, G) used in Perler et al. (1997). Also shown are the residues involved in the mechanism of canonical splicing. N-extein refers to the region of host precursor protein immediately upstream from the intein; C-extein refers to the region of host precursor protein immediately downstream from the intein. The double lines indicate the area in which a homing endonuclease is found in full-length inteins. In mini-inteins, only a short linker region is found in this region. (B) Representation of a full-length (homing endonuclease containing) PRP8 intein from *N. fischeri*. The four conserved motifs of the homing endonuclease domain are identified with the nomenclature of Perler et al. (1997). (C) Representation of a PRP8 mini-intein from *N. glabra*. Both PRP8 inteins are drawn to the same scale; data are from Butler et al. (2006).