Inteins invading mycobacterial RecA proteins

Isabelle Saves[a,*], Marie-Antoinette Lanéelle[a], Mamadou Daffé[a], Jean-Michel Masson[a,b]

[a] *Institut de Pharmacologie et Biologie Structurale (UMR5089), C.N.R.S./Université Paul Sabatier Toulouse III, 205 Route de Narbonne, F-31077 Toulouse Cedex, France*

[b] *Institut National des Sciences Appliquées, Complexe Scientifique de Rangueil, F-31077 Toulouse Cedex, France*

Received 14 July 2000; revised 27 July 2000; accepted 31 July 2000

Edited by Gianni Cesareni

Abstract Five new inteins were discovered in a survey of 39 mycobacterial strains that was undertaken to clarify the role of RecA inteins in mycobacteria. They are all inserted at the RecA-b site of the *recA* gene of *Mycobacterium chitae*, *M. fallax*, *M. gastri*, *M. shimodei* and *M. thermoresistibile* and belong to the *Mle*RecA allelic family. Sequence analysis showed that although only *M. tuberculosis* harbours an intein at the RecA-a site the sequence of the RecA-b site is well conserved between species. Furthermore, the presence of inteins does not correlate with specific characteristics of the species such as pathogenicity or growth rate.

Key words: Intein; Protein splicing; Mycobacterium; RecA

1. Introduction

Inteins are protein introns, post-translationally spliced from a large polypeptide precursor to yield the mature host protein. In most cases, inteins are inserted in functionally important domains of proteins which are essential for cell survival [1]. Therefore, splicing of the intein is absolutely necessary to ensure the host protein cellular role. The splicing pathway involves one residue of the host protein located immediately downstream of the intein, and residues belonging to four intein conserved motifs. Hence, inteins are positioned immediately upstream of a cysteine, a serine or a threonine residue required to achieve splicing. No essential function has yet been ascribed to inteins. Although the majority of known inteins contain conserved LAGLIDADG motifs of the homing endonuclease DOD family [2], this endonuclease activity seems only to be involved in the perpetuation and transfer of the intein [3] and has been experimentally demonstrated for only a few inteins. Intein coding sequences are widely distributed in all three kingdoms [4,5]. Among the 12 reported eubacteria containing inteins, eight are mycobacteria, namely *Mycobacterium avium*, *M. flavescens* (*Mfl*), *M. gordonae*, *M. kansasii*, *M. leprae* (*Mle*), *M. malmoense*, *M. tuberculosis* (*Mtu*) and *M. xenopi*.

The RecA recombinase of *M. tuberculosis* is required for homologous DNA pairing and strand exchange [6,7]. This protein is the product of protein splicing since its gene contains an in-frame insertion of 1320 bp which encodes an intein, at the RecA-a site [8,9]. In 1994, Davis et al. [10] have shown that *M. leprae* also harbours an intein in the *recA* gene but at a different insertion site (RecA-b), while 14 other mycobacterial *recA* genes examined do not have invading sequences. The occurrence of these inteins in these two obligate mycobacterial pathogens suggested that RecA inteins might play a role in mycobacterial functions related to pathogenesis or virulence. Alternatively, these inteins may be responsible for the low efficiency of homologous recombination in *M. tuberculosis* and other slowly growing mycobacteria.

To further gain insight into the potential roles of inteins through the distribution of inteins in the RecA proteins of mycobacteria, the present study of a vast range of mycobacterial species and strains was undertaken. For this purpose, we investigated 39 mycobacterial strains belonging to 32 different species. PCR amplification of *recA* genes from the genomic DNA of these mycobacteria revealed six inteins, which were cloned and sequenced.

2. Materials and methods

2.1. Mycobacterial strains and growth conditions

The 39 mycobacterial strains listed in Table 1 were from the Reference Centre for Mycobacteria (Institut Pasteur, Paris) and were kindly provided by Dr V. Vincent Lévy-Frébault. All the strains were identified using the conventional biochemical tests as described by Wayne and Kubica [11]. In addition, chemotaxonomic criteria such as mycolic acid and species-specific glycolipid analyses [12] were applied to all the strains examined. Strains were grown at their optimal temperature of growth [11] on Löwenstein–Jensen medium for a few days to several weeks, depending on the growth rate of the mycobacterial species.

2.2. Genomic DNA isolation

The glass bead disruption method for genomic DNA isolation from mycobacteria was chosen, based on the results obtained by Via and Falkinham [13] showing that this method is suitable for DNA amplification. The mycobacteria were scraped from Löwenstein–Jensen solid medium and resuspended in 1 ml of sterile TE buffer (Tris–HCl 10 mM pH 8, EDTA 1 mM). One gram of 0.1 mm glass beads (Bioblock) was added and the mixtures were vortexed for 2 min to disrupt the mycobacteria. Proteinase K (Sigma) was added at a final concentration of 50 μg/ml and the mycobacterial suspensions were incubated for 10 min at 65°C then for 10 min at 95°C. These lysates were centrifuged for 10 min at $5000 \times g$ to pellet debris and glass beads and the DNA-containing supernatants were treated first with phenol/chloroform/isoamyl alcohol (25/28/1) and then with chloroform. These DNA preparations were directly used in PCR assays.

2.3. PCR assays

Primer sequences were designed according to gene sequence conser-

*Corresponding author. Fax: (33)-5-61 17 59 94.
E-mail: saves@ipbs.fr

Abbreviations: *Mch*, *Mycobacterium chitae*; *Mfa*, *M. fallax*; *Mfl*, *M. flavescens*; *Mga*, *M. gastri*; *Mle*, *M. leprae*; *Msh*, *M. shimodei*; *Msmeg*, *M. smegmatis*; *Mth*, *M. thermoresistibile*; *Mtu*, *M. tuberculosis*

0014-5793/00/$20.00
PII: S0014-5793(00)01944-X

vation between mycobacterial species. The *recA* gene sequences from *M. leprae*, *M. tuberculosis* and *M. smegmatis* (*Msmeg*) were aligned using Multalin software [14]. Oligonucleotide sequences were chosen in the most conserved part of the gene surrounding both intein insertion sites. The primers RecA-5′ (5′-accacggcgatcttcatcaaccagct-3′) and RecA-3′ (5′-aggatgtcgaactcggccagcttgaa-3′) correspond to the *M. tuberculosis recA* gene sequence or its complementary sequence between positions 562–587 and positions 765–791, respectively.

PCR amplifications were performed using 5 μl of genomic DNA preparation as DNA matrix. Several PCR assays were performed using each oligonucleotide pair in different reaction conditions with regard to oligonucleotide concentration and hybridisation temperature to ensure amplification specificity. In the absence of invading sequence, the amplified fragment consisted of 229 bp of the *recA* gene.

2.4. Cloning and sequencing of intein coding sequences

PCR fragments containing invading sequences were purified by the Qiaquick extraction kit (Qiagen) and directly cloned in the PCR2.1-TOPO plasmid using the TOPO-TA cloning kit (Invitrogen). The resulting plasmids were double-strand sequenced (MWG-Biotech) using universal primers M13Forward and M13Reverse.

2.5. Accession numbers for nucleotide and peptide sequences

GenBank accession numbers for *Mch*, *Mfa*, *Mfl*, *Mga*, *Mth*, *Msh*, *Mle*, *Mtu* and *Msmeg recA* genes are AJ251336, AJ251337, AJ251150, AJ251715, AJ251714, AJ251338, X73822, X58485 and Q59560, respectively.

3. Results and discussion

Since the description of protein introns in RecA of both pathogens *M. leprae* and *M. tuberculosis* by Davies et al. [9,10], the possible role of these inteins in mycobacterial pathogenicity has been postulated but not demonstrated. In order to definitively assess the phylogenetic or physiological relevance of RecA inteins in mycobacteria, 39 mycobacterial strains belonging to 32 species (Table 1), 21 non-pathogenic and 11 opportunistic pathogens [11], were tested for the presence of intein in *recA* gene.

PCR amplifications of the genomic DNA of these strains were performed with oligonucleotides allowing the detection of an invading sequence in a 229 bp fragment of the *recA* gene containing both RecA-a and RecA-b insertion sites. The presence of an intein can thus be revealed by the specific amplification of a longer DNA fragment. Among the 39 strains tested, eight showed an invading sequence since a fragment of approximately 1300 bp was amplified from the genomic DNA of these strains, namely the three strains of *M. chitae* (*Mch*), and one strain of *M. fallax* (*Mfa*), *M. flavescens*, *M. gastri* (*Mga*), *M. shimodei* (*Msh*) and *M. thermoresistibile*

Table 1
A survey of inteins present in mycobacterial *recA* genes

Species	Strain reference	Intein presence and size	Insertion site
M. abcessus	IP 810402	–	
M. agri	ATCC 27406^T	–	
M. aichiense	IP 141270002	–	
M. aurum	IP 141210005	–	
M. aurum	ATCC 23366^T	–	
M. avium	IP140310013	–	
M. avium	serovar 8	–	
M. chitae	IP 141160001	+1092 (364)[a]	RecA-b
M. chitae	IP 141160002	+1092 (364)[a]	RecA-b
M. chitae	IP 141160003	+1092 (364)	RecA-b
M. chelonae	NCTC 946^T	–	
M. cookii	ATCC 49103	–	
M. duvalii	IP 141180001	–	
M. duvalii	IP 141180002	–	
M. duvalii	IP 141180003	–	
M. duvalii	IP 141180004^T	–	
M. fallax	IP 8139^T	+1089 (363)	RecA-b
M. farcinogenes	IP 141100002^T	–	
M. flavescens	ATCC 14474^T	+1092 (364)	RecA-b
M. fortuitum	ATCC 6841^T	–	
M. gadium	ATCC 27726^T	–	
M. gastri	HB 4389	+1104 (368)	RecA-b
M. gordonae	ATCC 14470^T	–	
M. kansasii	ATCC 12478^T	–	
M. marinum	IP 14 012 001^T	–	
M. nonchromogenicum	ATCC 19530^T	–	
M. parafortuitum	IP 141220001^T	–	
M. peregrinum	ATCC 14467^T	–	
M. phlei	ATCC 11758^T	–	
M. porcinum	IP 141460001^T	–	
M. rhodesiae	IP 141240001^T	–	
M. scrofulaceum	IP 140220023	–	
M. senegalense	ATCC 35796^T	–	
M. shimodei	ATCC 27962^T	+1092 (364)	RecA-b
M. simiae	IP 102 0017	–	
M. smegmatis	ATCC 607	–	
M. terrae	Portaels 4977	–	
M. thermoresistibile	ATCC 19527^T	+1095 (365)	RecA-b
M. triviale	IP140330002	–	

The exact size in bp (in amino acids) of inteins is indicated. – means that no invading sequence was detected. ^{T} means type strain.
[a]The intein was not cloned. The intein size was hypothesised from the sequence of the third *M. chitae* intein.

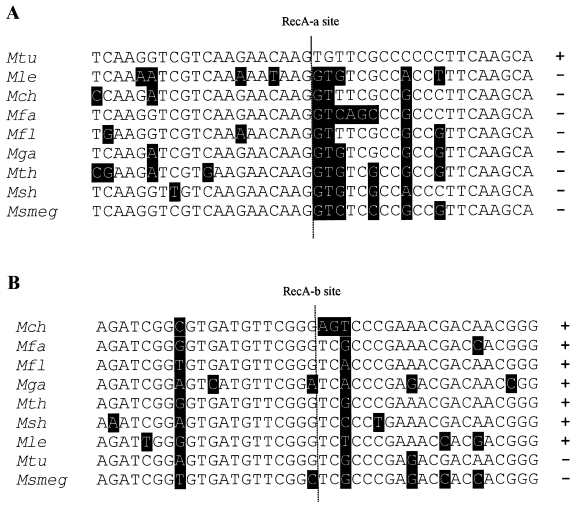

Fig. 1. Nucleotide sequence alignments of the RecA-a (A) and RecA-b (B) intein insertion sites in *recA* genes from different mycobacterial species. The + or −, at the right of the sequences, means that an intein is present or absent, respectively, at this insertion site. In A, nucleotides divergent from *recA* of *M. tuberculosis* appear in black boxes. In B, nucleotides divergent from the most commonly found nucleotides appear in black boxes.

(*Mth*). The exact size of these invading sequences and their insertion site were determined by cloning and sequencing the corresponding *recA* gene fragments (Table 1). These results confirmed the presence of an intein in *M. flavescens recA* at the same insertion site as *M. leprae* intein, which was suggested but not published by Adam in 1997 [5]. More importantly, they revealed five new homologous inteins at the same site in *recA* genes from of *M. chitae, M. fallax, M. gastri, M. shimodei* and *M. thermoresistibile*. According to the current nomenclature, these inteins were named *Mch*RecA, *Mfa*RecA, *Mga*RecA, *Msh*RecA and *Mth*RecA, respectively. These five inteins, together with that from *M. flavescens* (*Mfl*RecA), are characterised by a great similarity to their *M. leprae* equivalent, in terms of size, peptide sequence and location in the host protein. Thus, no new intein insertion site in RecA and no new mini-intein were discovered. Hence, among the 38 mycobacterial species tested to date ([10] and this work), eight possess an intein in RecA. RecA inteins were found in both pathogenic (*Mtu* and *Mle*) and non-pathogenic (*Mch*, *Mfa*, *Mfl*, *Mga*, *Msh* and *Mth*) strains. Additionally, the eight strains represent four slowly growing and four rapidly growing mycobacterial species. Hence, we demonstrated that, contrary to what has been hypothesised earlier, the presence of protein intron in RecA is not correlated with physiological properties of mycobacteria such as pathogenicity or growth rate.

Inteins *Mtu*RecA and *Mle*RecA are inserted at two distinct insertion sites, RecA-a and RecA-b, respectively. The present work shows that the six other RecA inteins are located in the RecA-b site. Hence, it is surprising that only the *M. tuberculosis* intein is inserted at the RecA-a site while seven inteins are now identified at the RecA-b site (Table 1). These results confirm and extend the study of Davis et al. [10] who did not find any equivalent to *Mtu*RecA intein while examining 14 different mycobacterial species by Southern hybridisation and by PCR. The RecA-b site is located immediately downstream of glycine 205 in *M. leprae*. The equivalent residue G204 in *Escherichia coli* RecA belongs to the unstructured loop L2 in the central conserved region of bacterial RecA proteins. Three-dimensional structure of *E. coli* RecA and functional analyses showed that this loop is involved in DNA binding [15]. Hence, its functional importance could explain the prevalence of inteins at this site. However, the RecA-a site also belongs to a highly conserved region of RecA proteins. Lysine 251, at the N-terminal junction between RecA and the intein, belongs to a loop for which no direct implication in RecA function has been determined, although this region of the protein may be indirectly involved in ATP binding [15]. Thus, the singularity of *Mtu*RecA intein insertion at this site cannot be explained by a functional analysis of RecA. Actually, even if the RecA peptide sequence is highly conserved among mycobacteria since RecA proteins of *Mle*, *Msmeg* and *Mtu* differ only by a few point substitutions from the RecA partial sequences of *Mch*, *Mfa*, *Mfl*, *Mga*, *Msh* and *Mth* (data not shown), the singularity of the *Mtu*RecA intein location is due to the divergence of the RecA peptide sequence at this site. Effectively, the comparison of the *recA* genes from *Mch*, *Mfa*, *Mfl*, *Mga*, *Mle*, *Msh*, *Msmeg*, *Mth* and *Mtu* at the RecA-a site highlights a major divergence (Fig. 1A). While the first codon at the 3′ side of the homing site of *Mtu*RecA encodes the cysteine required for splicing, the corresponding codon in the other mycobacterial *recA* genes encodes a valine. Hence, the insertion of an intein at this site is not possible, the splicing process being hindered by the valine at this position. That this reasoning also applies to RecA-b site was investigated by analysing the nucleotide sequences at this intein homing site. Alignment of the RecA-b insertion sites of the seven inteins of the *Mle*RecA allelic family with *recA* genes from *M. tuberculosis* and *M. smegmatis*, which are inteinless at this site, does not reveal crucial substitutions in nucleotide sequences (Fig. 1B). Moreover, the most divergent RecA-b site is that of *Mch*RecA in which the first codon downstream of the homing site, encoding the first serine of the C-extein, is AGT instead of a TCN codon in other known *recA* genes. Therefore, in contrast to the RecA-a site, a divergence in nucleotide sequence could not explain the absence of intein at this site of *Mtu* and *Msmeg recA* genes.

Table 2
Percentage identity between mycobacterial RecA intein peptide sequences

	*Msh*RecA	*Mth*RecA	*Mga*RecA	*Mfl*RecA	*Mfa*RecA	*Mch*RecA	*Mle*RecA
*Mle*RecA	86.6	79.0	65.0	83.3	81.1	81.1	100
*Mch*RecA	85.7	84.7	68.5	89.0	84.3	100	
*Mfa*RecA	84.1	81.6	66.3	83.2	100		
*Mfl*RecA	87.1	85.8	67.7	100			
*Mga*RecA	65.9	65.9	100				
*Mth*RecA	83.6	100					
*Msh*RecA	100						

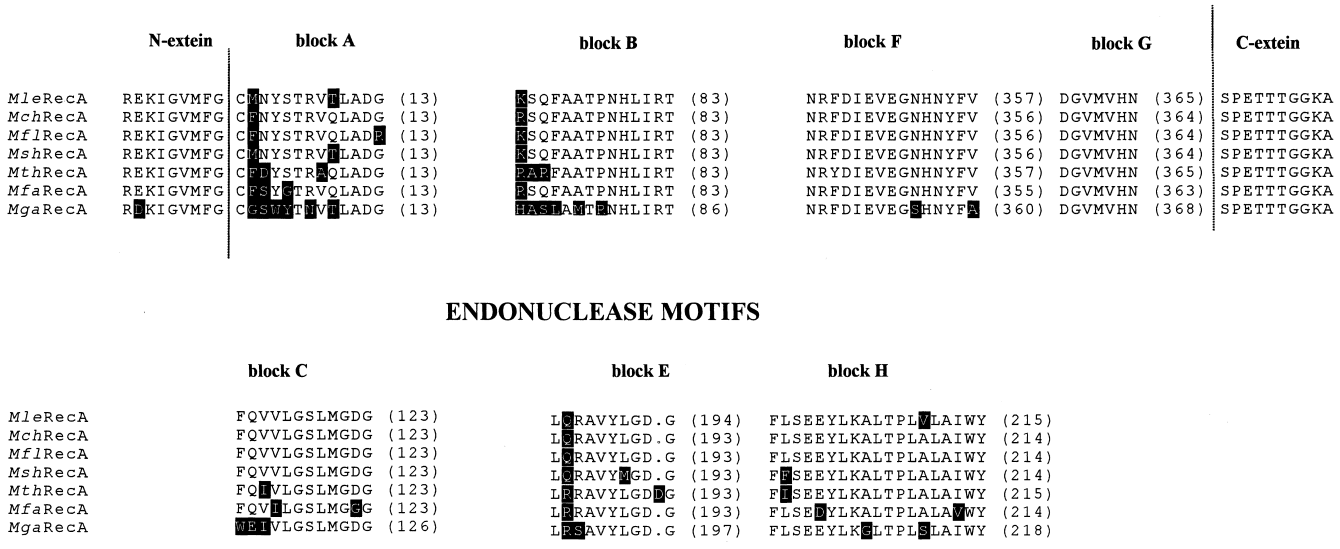

Fig. 2. Sequence alignment of the conserved motifs of the seven inteins of the *Mle*RecA allelic family. The sequences of N- and C-extein sequences are also indicated. The position of the last residue of each motif in the intein sequence appears between parentheses. The variable residues within these motifs and within N- and C-exteins appear in black boxes.

When the seven peptide sequences of the N- and C-terminal junctions between RecA and the intein as well as the seven sequences of inteins belonging to the *Mle*RecA allelic family were compared to those of *M. leprae*, the extein peptide sequences at the intein insertion sites were found to be highly conserved (Fig. 2), only one glutamic acid residue being replaced by an aspartic acid in the *M. gastri* RecA protein. In the same way, the seven inteins of the same allelic family share high sequence conservation since the peptide sequence of the newly described inteins are 65–87% identical to *Mle*RecA intein (Table 2). The degree of conservation between inteins fluctuates and it was not possible to find any correlation between sequence conservation and some phylogenetic similarities between invaded species.

Fifty-four per cent of the residues are conserved within the seven inteins. This high conservation allowed us to easily locate the intein motifs responsible for splicing (blocks A, B, F and G) and for endonuclease activity (blocks C, E and H) in these inteins. Except for the presence of a block D which is uncertain, the intein signature is well conserved among the seven inteins (Fig. 2). Even if the *Mga*RecA intein has the most divergent sequence (Table 2), the seven blocks are well conserved and the percentage identity with the other inteins is higher within the intein motifs than along the whole intein sequence. The high conservation between the RecA inteins from different species suggests that the seven inteins may share similar functions. Splicing of *Mle*RecA does not occur in *E. coli* [10] but produces mature RecA protein in *M. smegmatis* [16]. It would thus be interesting to check if sequence divergence in motifs A and B could affect this splicing. It seems plausible that the seven RecA inteins are specific endonucleases, since they possess the LAGLIDADG motifs of DOD endonucleases. However, deviations in block E of *Mth*RecA and in block C of *Mfa*RecA are particularly surprising. Indeed, in the second LAGLIDADG motif of *Mth*RecA, the conserved aspartic acid is duplicated, and the very conserved GDG triad is thus replaced by GDDG. In the case of *Mfa*RecA block C, the conserved aspartic acid is replaced by a glycine, which has not been described at this position to date [5]. One can expect that such variations in DOD motifs might play an important role in the specificity and efficiency of the putative endonuclease activity. Moreover, although the RecA-b insertion site sequences are well conserved (Fig. 1A), small divergences of these sequences may affect the endonuclease activity. It would thus be interesting to investigate the recognition of the seven target sequences by the seven inteins in order to determine the effect of the substitutions on specific recognition and cleavage.

References

[1] Dalgaard, J.Z., Klar, A.J., Moser, M.J., Holley, W.R., Chatterjee, A. and Saira Mian, I. (1997) Nucleic Acids Res. 25, 4626–4638.
[2] Mueller, J.E., Bryk, M., Loizos, N. and Belfort, M. (1993) in: Nucleases, 2nd edn. (Linn, S.M., Lloyd, R.S. and Roberts, R.J., Eds.), pp.111–143, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
[3] Gimble, F.S. and Thorner, J. (1992) Nature 357, 301–306.
[4] Perler, F.B., Olsen, G.J. and Adam, E. (1997) Nucleic Acids Res. 25, 1087–1093.
[5] Inbase, the New England Biolabs intein database. Page updated 04/12/2000. http://www.neb.com/neb/inteins/ [07/10/00, last date accessed].
[6] Kumar, R.A., Vaze, M.B., Chandra, N.R., Vijayan, M. and Muniyappa, K. (1996) Biochemistry 35, 1793–1802.
[7] Vaze, M.B. and Muniyappa, K. (1999) Biochemistry 38, 3175–3186.
[8] Inteins-Protein introns web site. Page updated in January 2000. http://bioinfo.weizmann.ac.il/–pietro/inteins/ [07/10/00, last date accessed].
[9] Davis, E.O., Sedgwick, S.G. and Colston, M.J. (1991) J. Bacteriol. 173, 5653–5662.
[10] Davis, E.O., Thangaraj, H.S., Brooks, P.C. and Colston, M.J. (1994) EMBO J. 13, 699–703.
[11] Wayne, L.G. and Kubica, G.P. (1986) In: Bergey's Manual of Systematic Bacteriology (Sneath, P.H.A., Mair, N.S., Sharpe, M.E. and Holt, J.G., Eds.), Vol. 2, pp. 1436–1457, Williams and Wilkins, Baltimore, MD.
[12] Daffé, M. and Draper, P. (1998) Adv. Microb. Physiol. 39, 131–203.

[13] Via, L.E. and Falkinham, J.O. (1995) J. Microbiol. Methods 21, 151–161.
[14] Multalin web site. Multiple sequence alignment by Florence Corpet. Page updated 03/28/2000. http://www.toulouse.inra.fr/multalin.html [07/10/00, last date accessed].
[15] Story, R.M., Weber, I.T. and Steitz, T.A. (1992) Nature 355, 318–325.
[16] Frischkorn, K., Springer, B., Bottger, E.C., Davis, E.O., Colston, M.J. and Sander, P. (2000) J. Bacteriol. 182, 3590–3592.