The EMBO Journal vol.13 no.3 pp.699-703, 1994

Evidence of selection for protein introns in the RecAs of pathogenic mycobacteria

Elaine O.Davis, Harry S.Thangaraj, Patricia C.Brooks and M.Joseph Colston

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

Communicated by J.J.Skehel

Protein introns are recently discovered genetic elements whose intervening sequences are removed from a precursor protein by an unusual protein splicing reaction. This involves the excision of a central spacer molecule, the protein intron, and the religation of the amino- and carboxy-terminal fragments of the precursor. The *recA* gene of *Mycobacterium tuberculosis* contains one such element and we now show that the other major mycobacterial pathogen, *Mycobacterium leprae*, also possesses a protein intron in its *recA*, although other mycobacterial *recA* genes do not. However, these two protein introns are different in size, sequence and location of insertion of their coding sequences into the *recA*s of *M.tuberculosis* and *M.leprae*, indicating that acquisition of the protein introns has occurred independently in the two species, and thus suggesting that there has been selection for splicing in the maturation of RecA in the pathogenic mycobacteria. The *M.leprae* protein intron provides an example of conditional protein splicing, splicing occurring in *M.leprae* itself but not when expressed in *Escherichia coli*, unlike most previously described protein introns. These observations suggest that protein introns may perform a function for their host, rather than being just selfish elements.

Key words: leprosy/mycobacteria/protein intron/protein splicing/tuberculosis

Introduction

Genes containing insertion elements which are removed not from the RNA but from a precursor protein have recently been identified. The first report described a subunit of the vacuolar ATPase of the yeast *Saccharomyces cerevisiae* (Hirata *et al.*, 1990; Kane *et al.*, 1990). Subsequently the recombination gene *recA* of *Mycobacterium tuberculosis* (Davis *et al.*, 1991, 1992) and a DNA polymerase gene of *Thermococcus litoralis* (Hodges *et al.*, 1992; Perler *et al.*, 1992) have also been shown to contain 'protein introns', the latter containing two such sequences. The most recent protein intron to be identified is again in a vacuolar ATPase, in the pathogenic yeast *Candida tropicalis* (Gu *et al.*, 1993). In each case homology to examples of the equivalent gene from other organisms is disrupted by an unrelated sequence, the protein intron.

Evidence that splicing occurs at the protein level rather than the RNA level includes (i) pulse-chase analysis demonstrating conversion of a large precursor protein to the two normal products using a slow processing mutant, (ii) the requirement for translation to proceed right through the protein intron and (iii) the observations that mutations at splice sites which alter amino acid sequence inhibit the process whereas those which change only nucleic acid sequence do not (for reviews see Shub and Goodrich-Blair, 1992; Wallace, 1993). Studies using the yeast system have confirmed that a new peptide bond is formed between the two halves of the spliced protein (Cooper *et al.*, 1993). The process is expected to be autocatalytic based on the range of heterologous systems where splicing still occurs.

In at least two of the known examples the excised protein intron has endonuclease activity, recognizing and cutting at the DNA sequence into which its coding DNA is normally inserted (Gimble and Thorner, 1992; Perler *et al.*, 1992) and the yeast example will insert into an intronless allele in a homing process (Gimble and Thorner, 1992) analogous to that of group I mobile RNA introns (Perlman and Butow, 1989). Although the homology between these protein introns is not great, some motifs can be discerned. In each case there is an identifiable pair of dodecamer sequences bearing similarity to the LAGLIDADG motif at similar spacing to those found in some other endonucleases, especially the homing endonucleases encoded by some group I RNA introns (Doolittle, 1993), suggesting that the endonuclease function may be a common feature. In addition, there are conserved features around the splice sites with a cysteine or serine residue at the first splice site and a hexapeptide motif at the second splice site (Davis *et al.*, 1991; Hodges *et al.*, 1992; see Results).

Following our observation with *M.tuberculosis*, the causative agent of tuberculosis in man, we have now investigated the *recA* genes of other members of the genus. Many species of mycobacteria are harmless environmental organisms, although some are opportunistic pathogens or pathogens for animals. In addition to *M.tuberculosis* the other mycobacterial species which is a natural pathogen of man is *Mycobacterium leprae* which causes leprosy. Here we report that the *recA* gene of *M.leprae* also contains a protein intron, while the *recA* genes of other mycobacteria do not. Furthermore, these protein introns appear to have been acquired independently by the two pathogenic species. Based on our observations we propose that protein introns may have a role to play in their bacterial hosts' survival rather than being just selfish elements.

Results

Cloning the M.leprae recA gene

Following our observation with the *recA* gene of *M.tuberculosis*, we investigated the presence of similar protein intron sequences in other mycobacteria by Southern hybridization using intron-encoding DNA; only organisms belonging to the *M.tuberculosis* complex gave positive

hybridization (data not shown), suggesting that protein splicing was not a common feature of mycobacterial RecAs. We then cloned the *M.leprae recA* gene from cosmid clones of the *recA* region of the *M.leprae* genome (Eiglmeier *et al.*, 1993) after analysis by hybridization with probes from each end of the *M.tuberculosis recA* gene to identify a restriction fragment containing the whole gene. A single fragment hybridizing to both probes of a suitable size for cloning was found for only two of 24 enzymes tested, and one of these, a 6.5kb *Sph*I fragment, was cloned into pTZ18R to give pEJ216.

Further mapping and hybridization to this subclone identified the locations at which the two probes bound from restriction sites which they spanned. In particular both probes bound to a common 1.8 kb *Msc*I fragment in addition to a second fragment which was different for each probe, indicating that the two probes bound ~1.8 kb apart, such a distance that an intervening sequence must be present in the *M.leprae recA* gene even though the *M.tuberculosis* intron did not hybridize (Figure 1). It was confirmed that no rearrangement had occurred during the cloning procedure by hybridization to genomic *M.leprae* DNA digested with four different enzymes using this fragment as a probe, revealing bands seen in this clone or the original cosmid (data not shown).

The M.leprae RecA contains a protein intron

The analysis described above indicated that the whole *recA* gene should be contained on a 3.3 kb *Sph*I–*Hpa*I fragment so this was further subcloned into pUC19 to give pEJ217. Tn*1000* insertions were isolated in this cloned fragment and used for sequencing with primers to the two ends of the transposon. This sequencing of the *M.leprae recA* gene confirmed the presence of an intervening, protein intron-like sequence, with homology to other *recA* genes being split by a stretch of unrelated sequence, and the hexapeptide motif previously identified at the carboxy-terminus of protein introns (Davis *et al.*, 1991; Hodges *et al.*, 1992) being quite well conserved (Figure 2A). However, the predicted size of the *M.leprae* intron was 41 kDa, compared with 47 kDa for *M.tuberculosis*. Moreover, apart from the carboxy-terminal motif, the sequences of the two protein introns were quite dissimilar (only 27% identity, Figure 2B), whereas the RecA-like sequences were very homologous with 92% of amino acids being identical (data not shown, EMBL/GenBank accession number X73822). Most surprising, however, was the fact that the two protein introns were located at different insertion points within the RecA coding sequences in the two species (Figure 2C). This is in contrast to the situation with the protein introns in the vacuolar ATPase of the yeasts *S.cerevisiae* and *C.tropicalis* which are each located at an identical position in the host gene. The nucleotide sequences into which the protein introns are inserted in the two mycobacteria are also quite different (Figure 2C), although the *M.leprae* sequence has significant similarity to the equivalent site of *S.cerevisiae* (highlighted in Figure 2C). Thus it appears that insertion of the protein intron sequences into the two species must have occurred independently.

Other mycobacterial recA genes do not contain protein introns

The discovery of a second mycobacterial *recA* gene containing a protein intron was surprising since the

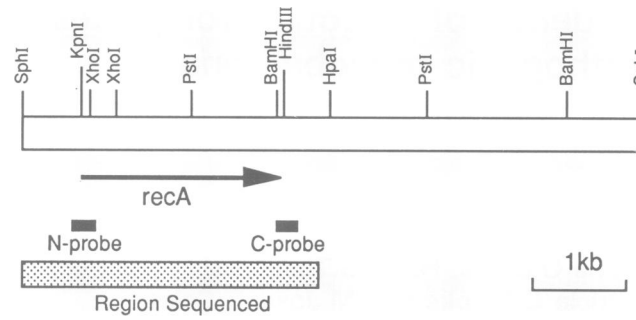

Fig. 1. Restriction map of the 6.5 kb *Sph*I fragment cloned from cosmid clones of the *M.leprae* genome. The arrow represents the *recA* gene and its direction of transcription; the two small boxes show where the amino-terminal and carboxy-terminal *M.tuberculosis recA* probes bound, and the stippled box indicates the region sequenced.

M.tuberculosis intron did not hybridize to other mycobacterial genomes except for the very closely related *M.bovis* BCG and *M.microti* (which form part of the *M.tuberculosis* complex), and so the occurrence of such sequences in other mycobacterial *recA* genes was investigated. Primers based on sequences conserved between *M.leprae* and *M.tuberculosis* from near to the two ends of the *recA* gene were used in a PCR analysis, the size of the product revealing the presence (2.2 kb for the *M.tuberculosis* complex or 2.0 kb for *M.leprae*) or absence (0.9 kb) of an intron. A further 14 species of mycobacteria representative of various groups were examined (Figure 3) but none of these possessed an intervening sequence. In contrast, all clinical isolates of *M.tuberculosis* tested did contain an intron (data not shown). The PCR products obtained were confirmed to be *recA* by hybridization with a probe internal to RecA coding sequences. Sequencing of the PCR product obtained from *M.microti* showed that the protein intron was inserted in the same location and into the same sequence as in *M.tuberculosis*, and that the sequences at the two ends of the protein intron itself were also identical, in accord with these two species belonging to the *M.tuberculosis* complex. It would therefore appear that the presence of protein introns in mycobacterial RecAs is confined to the *M.tuberculosis* complex and *M.leprae*.

Conditional splicing of the M.leprae protein intron

Expression of the *M.leprae recA* gene from a T7 promoter and labelling of its products with [^{35}S]methionine after inhibition of *Escherichia coli* RNA polymerase with rifampicin revealed a single large product corresponding to the entire open reading frame (Figure 4A). This is in contrast to the situation with the *M.tuberculosis recA* gene which is spliced to form two smaller proteins (shown for comparison) and is despite the conserved features at the two ends of the protein intron: a Cys (of Cys or Ser) at the amino-terminus and the motif mentioned above at the carboxy-terminus (Figure 2A). All protein introns identified to date share these features and have been found to splice at least in the native organism. Therefore, we examined the status of the RecA protein in extracts of *M.leprae* using antiserum raised against the *E.coli* RecA protein, and found only one band on a Western blot of approximately the same molecular mass as *E.coli* RecA (Figure 4B). Thus the RecA precursor is completely spliced in *M.leprae* itself. Attempts to identify conditions under which the *M.leprae* RecA would splice in

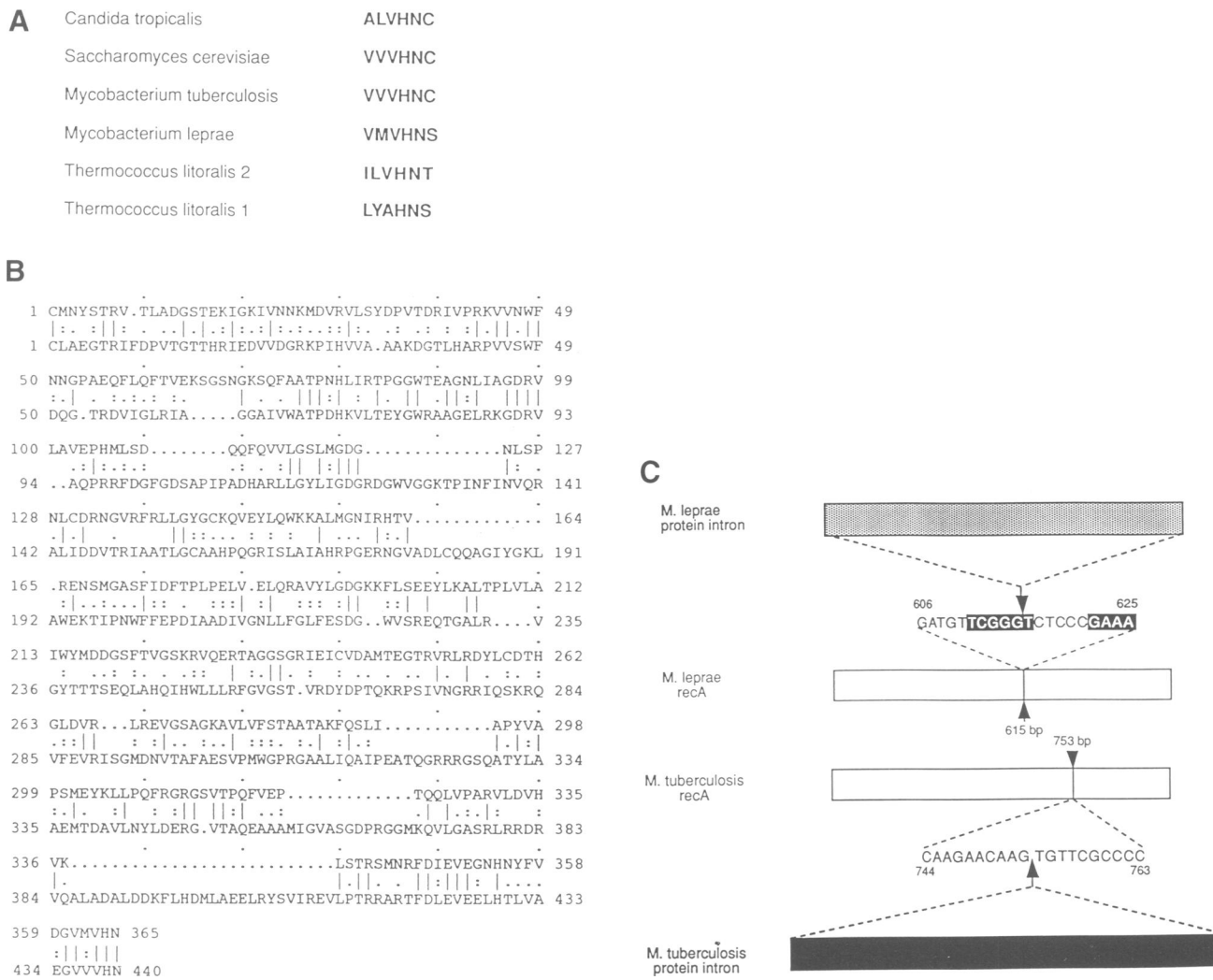

Fig. 2. Comparison of the *M.leprae* protein intron with that of *M.tuberculosis*. (**A**) Alignment of the conserved hexapeptide motif around the second splice site for the known examples of protein introns. The cut site is immediately before the C/S/T. (**B**) Comparison of the amino acid sequences of the protein introns of *M.leprae* (upper line) and *M.tuberculosis* (lower line) showing little overall homology, generated by the program GAP of the GCG package (Devereux *et al.*, 1984) using default parameters. (**C**) Schematic illustration of the structure of the *M.leprae recA* gene compared with that of *M.tuberculosis*. The sites of insertion of the protein introns at different locations in the two *recA* genes are indicated in the central part of the figure; the differently shaded boxes representing the protein introns indicate the lack of homology between them. The expanded portion shows the actual sequences into which the introns are inserted with the highlighted sequence in *M.leprae* being conserved at the protein intron insertion site in the *VMA1* gene of *S.cerevisiae*.

E.coli were carried out by changing pH, temperature, the oxidative state and ions; however, we were unable to detect splicing under any such conditions (data not shown).

Discussion

In this study we have identified a second protein intron in a mycobacterial *recA* gene, that of *M.leprae*. In common with other protein introns described previously it disrupts the coding sequence for the gene in which it is found by an in-frame insertion of unrelated sequence. The motif at the second splice site of protein introns which has been implicated in the protein splicing reaction (Davis *et al.*, 1992; Hirata and Anraku, 1992; Hodges *et al.*, 1992; Cooper *et al.*, 1993) is quite well conserved, as is the cysteine/serine residue at the first splice site. In addition, sequences similar to the LAGLIDADG motifs found in other protein introns (Shub and Goodrich-Blair, 1992) and some homing endonucleases of group I RNA introns (Doolittle, 1993) are identifiable at a similar spacing, so it is possible that the *M.leprae* protein intron has endonuclease activity as well.

However, despite these general similarities, the *M.leprae* protein intron is quite different from that of *M.tuberculosis*. It is different in size, sequence and location of insertion of its coding DNA into the *recA* gene. The sequences of the two insertion sites in the *recA* genes are also different. In the *M.leprae* gene the site equivalent to the insertion point in *M.tuberculosis* is substantially different at the cut site which would presumably prevent insertion of a tuberculosis-type element were it to enter the cell. Similarly, there are differences in sequence, albeit less dramatic, in the *M.tuberculosis* gene at the site of the *M.leprae* insertion. These differences between the introns in the two species suggest that they have been acquired independently. Given the rarity of such genetic elements, their presence in the same gene in the two species could not have arisen by chance, and suggests that there has been selection for them, and thus that protein introns have a function for the host organism.

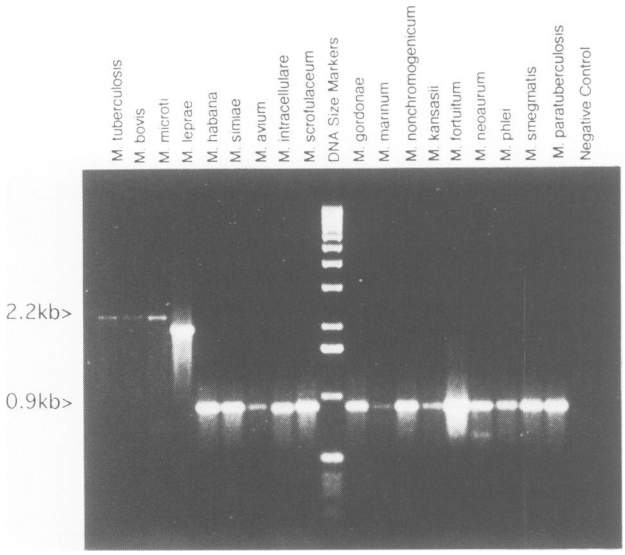

Fig. 3. The presence of a protein intron is confined to the *recA* genes of *M.leprae* and the *M.tuberculosis* complex. Primers to conserved sequences from near the ends of the *M.leprae* and *M.tuberculosis recA* genes were used in PCRs on genomic DNA from the various mycobacterial species indicated (*M.bovis* referring to *M.bovis* BCG). A large product (2.2 kb for the *M.tuberculosis* complex and 2.0 kb for *M.leprae*) reveals the presence of a protein intron, whereas the size of the product from the other species (0.9 kb) is that expected for a *recA* gene with no insertion.

Their presence in only the two species of mycobacteria which are major human pathogens raises the possibility of a role for protein introns in intracellular survival or pathogenesis.

The *M.leprae* protein intron was found to splice in *M.leprae* itself but not when expressed in *E.coli* under conditions where the *M.tuberculosis* protein intron did splice. In *T.litoralis* the first protein intron in the Vent DNA polymerase, termed IVS1, also failed to splice in *E.coli* (Perler *et al.*, 1992), although in this case it could not be cloned with the whole gene, so it is possible that the absence of parts of the polymerase and the second protein intron, IVS2, could have prevented folding into the correct structure required for splicing. As the protein introns examined in most detail seem to be able to splice out of a completely foreign context, this explanation seems unlikely, but it is possible that IVS2 is required for IVS1 to splice. In the case of the *M.leprae* protein intron it may be that some accessory mycobacterial protein is required to participate in splicing or that the activity is controlled in some way which might involve a regulatory protein. Such conditional splicing would provide a novel mechanism for regulation of RecA activity, as we have previously shown that the unspliced *M.tuberculosis* precursor lacks RecA activity (Davis *et al.*, 1992).

The *recA* gene is important for the repair of DNA damage which may be caused by oxidative stress, one of the conditions to which these intracellular pathogens are exposed when they invade macrophages. It is therefore likely that RecA would be important for survival within the cell and that protein introns may have a role in regulation of RecA expression under certain conditions. It is also possible that they might possess some other function more directly involved in pathogenesis or virulence. In any event the results described here suggest for the first time that protein introns have a role to play in their hosts' survival rather than being just selfish elements.

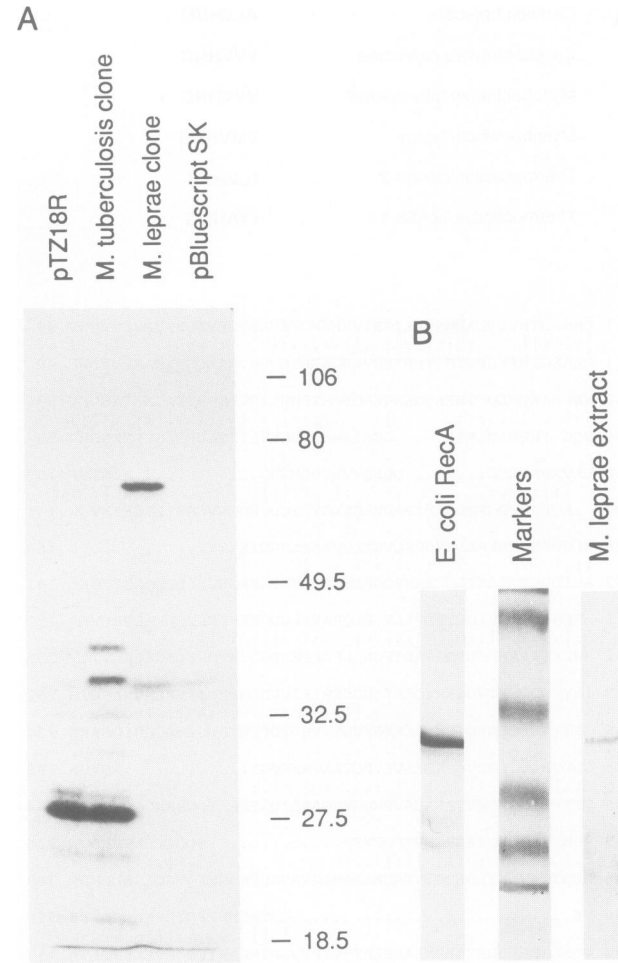

Fig. 4. Protein products of the *recA* locus from *M.leprae*. (**A**) Proteins expressed in *E.coli* from the T7 promoter were labelled with [^{35}S]methionine after inhibition of *E.coli* RNA polymerase with rifampicin. The single, high molecular mass, unspliced product from the *M.leprae* clone is compared with the two smaller spliced products from the *M.tuberculosis* clone; lanes are as indicated and the positions and sizes of molecular mass markers are given. (**B**) Western blot of *M.leprae* extract with antiserum raised against *E.coli* RecA protein, compared with purified *E.coli* RecA. The lanes are as indicated and the sizes of the molecular mass markers are 80, 49.5, 32.5 and 27.5 kDa.

Materials and methods

DNA techniques

DNA manipulations were performed by standard methods (Sambrook *et al.*, 1989). Hybridizations were performed using the digoxigenin system (Boehringer) with washes in 2 × SSC/0.1% SDS at 60°C. The DNA fragment to be sequenced (3.3 kb *Sph*I−*Hpa*I) was subcloned in pUC19 (Yanisch-Perron *et al.*, 1985) to give pEJ217 and Tn*1000* insertions isolated in the cloned fragment. The sequence was obtained on both strands using primers from each end of the transposon (Thomas *et al.*, 1990) in double-stranded sequencing reactions using Sequenase (USB) by a modification of the supplier's method so the reactions could be performed in microtitre plates. The sequence has been deposited in the EMBL database, accession number X73822.

For PCR analysis of mycobacterial *recA* genes the primers were based on conserved regions at the ends of the *recA* genes of *M.tuberculosis* and *M.leprae*; their sequences were GGCAAAGGTTCGGTGATGCG and TCCTTGATCTTCTTCTCGATCTC. The PCR conditions were 94°C for 1 min, 59°C for 1 min, and 72°C for 1 min for 25 cycles.

Analysis of expressed proteins

For analysis of the protein products of the *M.leprae recA* gene in *E.coli* a 2.4 kb *Kpn*I−*Bsa*I fragment was subloned into pBluescript SK$^-$

(Stratagene) such that expression would be from the T7 promoter, yielding plasmid pEJ243. Strain BL21(DE3) (Studier *et al.*, 1990; a lysogen of BL21 containing a chromosomal gene for T7 RNA polymerase under control of the *lac*UV5 promoter) carrying pEJ243 or pEJ135 [Davis *et al.*, 1991; the *M.tuberculosis recA* cloned in pTZ18R (Pharmacia)] was grown in M9 minimal medium plus required supplements to an A_{600} of ~0.5 and then induced by addition of IPTG to 0.4 mM for 30 min; rifampicin was added to 200 μg/ml for 30 min, then [^{35}S]methionine (40 μCi/ml) was added for 10 min before harvesting. Samples equivalent to 100 μl of culture were run on a 12.5% SDS–polyacrylamide gel, which was fixed with 25% propan-2-ol/10% acetic acid, treated with Amplify (Amersham), dried and autoradiographed, the film being exposed for 2 h.

M.leprae cell free extract was prepared as described by Ibrahim *et al.* (1990). Western blotting was performed by standard techniques (Sambrook *et al.*, 1989).

Acknowledgements

We thank S.West of ICRF for the generous gift of anti-RecA (*E.coli*) antiserum, S.Cole of Institut Pasteur for providing cosmid clones of the *M.leprae recA* region, J.J.McFadden of the University of Surrey for supplying genomic DNA of *M.paratuberculosis* and S.G.Sedgwick and P.J.Jenner of the National Institute for Medical Research for helpful suggestions.

References

Cooper,A.A, Chen,Y.-J., Lindorfer,M.A. and Stevens,T.H. (1993) *EMBO J.*, **12**, 2575–2583.

Davis,E.O., Sedgwick,S.G. and Colston,M.J. (1991) *J. Bacteriol.*, **173**, 5653–5662.

Davis,E.O., Jenner,P.J., Brooks,P.C., Colston,M.J. and Sedgwick,S.G. (1992) *Cell*, **71**, 201–210.

Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.

Doolittle,R.F. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 5379–5381.

Eiglmeier,K., Honore,N., Woods,S.A., Caudron,B. and Cole,S.T. (1993) *Mol. Microbiol.*, **7**, 197–206.

Gimble,F.S. and Thorner,J. (1992) *Nature*, **357**, 301–306.

Gu,H.H., Xu,J., Gallagher,M. and Dean,G.E. (1993) *J. Biol. Chem.*, **268**, 7372–7381.

Hirata,R. and Anraku,Y. (1992) *Biochem. Biophys. Res. Commun.*, **188**, 40–47.

Hirata,R., Ohsumi,Y., Nakano,A., Kawasaki,H., Suzuki,K. and Anraku,Y. (1990) *J. Biol. Chem.*, **265**, 6726–6733.

Hodges,R.A., Perler,F.B., Noren,C.J. and Jack,W. (1992) *Nucleic Acids Res.*, **20**, 6153–6157.

Ibrahim,M.A., Lamb,F.I. and Colston,M.J. (1990) *Int. J. Leprosy*, **58**, 73–77.

Kane,P.M., Yamashiro,C.T., Wolczyk,D.F., Neff,N., Goebl,M. and Stevens,T.H. (1990) *Science*, **250**, 651–657.

Perler,F.B. *et al.* (1992) *Proc. Natl Acad. Sci. USA*, **89**, 5577–5581.

Perlman,P.S. and Butow,R.A. (1989) *Science*, **246**, 1106–1109.

Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning. A Laboratory Manual*. 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Shub,D.A. and Goodrich-Blair,H. (1992) *Cell*, **71**, 183–186.

Studier,F.W., Rosenberg,A.H., Dunn,J.J. and Dubendorff,J.W. (1990) *Methods Enzymol.*, **185**, 60–89.

Thomas,S.M., Crowne,H.M., Pidsley,S.C. and Sedgwick,S.G. (1990) *J. Bacteriol.*, **172**, 4979–4987.

Wallace,C.J.A. (1993) *Protein Sci.*, **2**, 697–705.

Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) *Gene*, **33**, 103–119.

Received on October 12, 1993