The nucleotide sequence of the tnpA gene completes the sequence of the Pseudomonas transposon Tn501

Nigel L. Brown, Joseph N. Winnie, David Fritzinger* and R. David Pridmore+

Department of Biochemistry, University of Bristol, University Walk, Bristol BS8 1TD, UK

Received 21 June 1985; Accepted 22 July 1985

ABSTRACT
The nucleotide sequence of the gene (tnpA) which codes for the transposase of transposon Tn501 has been determined. It contains an open reading frame for a polypeptide of M,=111,500, which terminates within the inverted repeat sequence of the transposon. The reading frame would be transcribed in the same direction as the mercury-resistance genes and the tnpR gene. The amino acid sequence predicted from this reading frame shows 32% identity with that of the transposase of the related transposon Tn3. The C-terminal regions of these two polypeptides show slightly greater homology than the N-terminal regions when conservative amino acid substitutions are considered. With this sequence determination, the nucleotide sequence of Tn501 is fully defined. The main features of the sequence are briefly presented.

INTRODUCTION
Tn501 (1), from Pseudomonas aeruginosa, is a member of the "Tn3 family" of transposons (2), in that it has inverted terminal repeats of 38 base pairs, which are partially homologous to those of Tn3; and it is flanked by five base-pair direct repeats generated from the recipient replicon during the transposition process (3). Several transposons, including Tn21, Tn1721 and Tn2603, are known to have transposition functions sufficiently related to those of Tn501 that complementation of mutants in the transposition genes can occur (4,5), and models for the evolutionary relationship between these transposons have been proposed (6-9). Several other transposons have been identified which appear to be closely related to Tn21 (9).

The major mechanism of transposition of the Tn3-related transposons involves two steps, each requiring a transposon-encoded gene product (10,11). The first step requires the product of the tnpA gene together with host-coded replication functions, and involves the formation of a co-integrate molecule between the donor replicon containing the transposon, and the recipient replicon. The co-integrate molecule contains two copies of the
transposon in direct repeat. The second event is mediated by the product of the tnpR gene, and is an intramolecular recombination between the two copies of the transposon in the cointegrate. These two steps require cis-acting components in addition to the gene products; the formation of the cointegrate is absolutely dependent on the presence of the inverted terminal repeats of the transposon, and the resolution of the cointegrate by transposition resolvase is by recombination between the internal resolution (res) sites in each copy of the transposon. Current models of replicative transposition (see 2) suggest that the transposase catalyses three steps in the formation of the cointegrate: site-specific cleavage in one strand at each terminus of the transposon, a less specific double-stranded staggered cleavage in the recipient replicon, and the ligation of the transposon termini with the recipient replicon. The determination of the nucleotide sequence of the tnpA gene and the concomitant prediction of the amino acid sequence of the transposase is an essential part of elucidating the molecular mechanism of transposition.

In this paper we present the sequence of nucleotides 5301 - 8355 of Tn501, which includes the coding sequence for the tnpA gene. We compare the predicted amino acid sequence of the gene product with that of the Tn3 transposase as a preliminary step in locating those regions of the transposase involved in recognition of DNA and in catalysis. This study completes the DNA sequence of Tn501, and the main features of the sequence of the transposon are briefly presented. Tn501 is finding use as a mutagen to locate and define genes in cyanobacteria (12), and it is one of the simplest members of a widely-dispersed class of mercury-resistance transposons (7,9) which are associated with multiple resistance to antimicrobial agents. The information summarised here will be of use to those working on the molecular biology of cyanobacteria, and to those interested in the evolution of transposons and of antimicrobial resistance determinants.

METHODS

DNA sequence analysis was carried out by the random cloning of fragments of pJOE114 DNA (a SalGI-EcoRI deletion of pBR322 carrying Tn501; 13,14) in M13mp7, and related bacteriophage (15,16), followed by chain-termination sequence analysis (17). The nucleotide sequence of the whole transposon was determined on both strands. When problems were encountered in the sequence analysis due to persistent secondary structure these were resolved by using formamide gels or deoxyinosine substitution as
Programs for the storage and analysis of DNA sequence data were those described elsewhere (18). The program used for the prediction of protein secondary structure was that described by McLachlan (22).

RESULTS

The DNA sequence of nucleotides 5301 to 8355 of transposon Tn501, containing the \textit{tnpA} gene and its flanking sequences, is shown in Fig. 1. Nucleotides 5301-5398 have been presented in a previous publication describing the \textit{tnpR} gene of Tn501 (23), and the inverted repeat sequence at nucleotides 8318-8355 has also been described earlier (3). The nucleotide sequence has a G+C content of 66\%, and contains one major open reading frame which would code for a polypeptide of 988 amino acids, $M_r = 111,500$.

We have presented data (23) showing that the termination codon for the \textit{tnpR} gene of Tn501 was that at positions 5350-5352 (Fig. 1). There are only three nucleotides between the third base of this codon and the first base of the suggested initiation codon of the \textit{tnpA} gene. We have no formal evidence that we have correctly identified the initiation codon of the \textit{tnpA} gene, but the predicted amino acid sequence of the gene product aligns well with the amino acid sequence determined for the transposase from transposon Tn3 (24,25), as shown in Fig. 2. The prospective Shine-Dalgarno sequence is the GGA at positions 5344-5346, which is within the coding sequence of the \textit{tnpR} gene. The nearest alternative initiation codon is a GTG at position 5512, which has no apparent Shine-Dalgarno sequence. The codon usage of the proposed \textit{tnpA} gene is similar to that described for other genes in Tn501 (14,23,26,27), there being a preference for codons containing C or G at the third position. There are several other open reading frames in the sequence presented here, the largest being on the complementary strand, starting at nucleotide 5922 and extending beyond the sequence in Fig. 1 to nucleotide 4660. These reading frames do not have the same predicted codon usage as other Tn501 genes, and are probably not functional. The relatively high GC content of the DNA gives a lower statistical probability of there being termination codon sequences present, thus giving rise to longer open reading frames.

There is a number of short inverted and directly repeated sequences within the DNA sequence presented here. This may be a mere statistical consequence of the high GC content of the DNA, although the inverted
Figure 1. DNA sequence of the tnpA gene of Tn501 showing the predicted primary sequence of the gene product. The predicted amino acid sequences of both the C-terminus of the tnpR gene product (nucleotides 5301-5349) and the complete tnpA gene product (nucleotides 5336-8319) are shown. Amino acid sequences are in the single letter code, with asterisks marking the termination codons. The postulated Shine-Dalgarno sequence (5343-5345) and the potential stem-loop structure discussed in the text (8310-8338) are underlined, and the 38bp terminal inverted repeat sequence is also marked.

Repeats may affect the rate of synthesis or the stability of mRNA in this region. However, there is one short-range inverted repeat that occurs at a noteworthy position, and may have a specific biological function. This lies between nucleotides 8310 and 8338, and consists of a 7bp inverted repeat separated by 15bp containing the termination codon for the tnpA gene and the inside end of the terminal inverted repeat of the transposon. The corresponding region of transposon Tn3 (24) also has the potential to form a stem-loop structure in approximately the same place, and this also contains the tnpA gene terminator and the inside end of the inverted terminal repeat. In Tn3 the stem would be of six nucleotides, and the loop would be 14 nucleotides. These sequences may form a stem-loop structures at the end of the tnpA gene transcript and be involved in transcription termination (28). Alternatively, or additionally, these stem-loop structures may be capable of the attenuation of transcripts initiating in the vector replicon which would otherwise transcribe the non-coding strand of the tnpA gene. There is no such stem-loop structure which could attenuate transcription at the other end of the transposon in either Tn501 or Tn3.

There is 32% amino acid identity between the transposases of Tn501 and Tn3 when the sequences are aligned by parsimony procedures. There is a number of conservative amino acid substitutions between the sequences, but these have not been scored in the alignment. These conservative substitutions are scored, however, in the DIAGON plot (Fig. 2b), which show that the homology between the C-terminal regions is slightly greater than that between the N-terminal regions. Certain oligopeptide sequences are strongly conserved between the transposases of Tn501 and Tn3 (Fig. 2a). The significance of these is not yet clear. The more hydrophobic regions of
Figure 2 Alignment of the predicted gene products of the \textit{tnpA} genes of Tn501 and Tn3 (24,25) by (a) parsimonious alignment of identical amino acid residues; and (b) dot matrix analysis (DIAGON; 20) in which conservative amino acid substitutions are scored. In (a) the upper sequence
is that from Tn501. Amino acid identities are marked with a colon, and the hyphens are padding characters to help alignment. In (b) a window of 15 residues was used for the DIAGON alignment, and a match was scored if the homology was greater than that expected to occur at a frequency of $10^{-4}$ for proteins of the same amino acid compositions.

the two polypeptides are similarly distributed along their respective primary structures, which is expected if the polypeptides adopt similar tertiary structures.

The data presented in this paper, together with the DNA sequences of the mercury resistance genes (14,26,27) and the res-tnpR region (23), complete the DNA sequence of transposon Tn501. The full sequence has been lodged with the EMBL Nucleotide Sequence Library via the Cambridge Nucleotide Sequence Data Library Service (29). The physico-genetic map of Tn501 (i.e. the gene map derived from the DNA sequence) is shown in Fig. 3, and the positions of gene boundaries and other features are given in Table 1. Data which are derived from the full sequence are only presented in the text if they are concerned with the transposition functions. As Tn501 is used for transposon mutagenesis (12), we give the location of some of the restriction endonuclease cleavage sites which occur less than five times in Tn501 (Table 2).

DISCUSSION

Tn501 transposase

The determination of the DNA sequence of the Tn501 tnpA gene presented here has allowed the prediction of the primary structure of a second transposase of the Tn3 family, the other being that of Tn3 itself (24,25). In the absence of further biological data it is difficult to assess the significance of the homologies and differences between the two transposases. The transposases from Tn501 and Tn3 have completely analogous functions in transposition of their parental transposons, but they will not complement one another, as determined by using TnpA\(^{-}\) mutants (4). Thus, each transposase must contain sites required for the specific recognition of the ends of the transposon, which differ in detail between the Tn501 and Tn3 enzymes; and catalytic sites for the cleavage-ligation reaction, which may be very similar in both.

We have tried to identify regions of transposase that are involved in DNA recognition by looking for primary sequences which could give rise to the helix-turn-helix tertiary structure common to several DNA-binding
Figure 3. Physico-genetic map of Tn501 showing the relative locations of the known genes and major open reading frames. The terminal inverted repeats and the major promoter (39) are marked. The gene designated merP is that described as merC in (26), and has been renamed because the merC gene originally identified in plasmid R100 by genetic criteria (40) corresponds to a reading frame that is not present in Tn501 (41). The reading frames urf-1 and urf-2 have not been ascribed a function. The exact positions of gene boundaries and other features are given in Table 1, as are references to the sequence data. The transposon is 8355 nucleotide pairs in length.

...proteins (30). This structural motif contains an 'invariant' glycine and hydrophobic residues four amino acids before and six amino acids after the glycine. The method of predicting secondary structure that was used (22) coupled with examination of the sequence for homology to these conserved residues failed to identify a strong candidate for such a DNA binding domain which occurs in both transposases. All methods of predicting secondary structure have weaknesses, and a helix-turn-helix structure may be present, but not have been detected. The best candidate for such a sequence is that around Gly-854 of the Tn501 transposase. An alternative explanation is...
Table 2. Restriction endonuclease cleavage sites occurring less than five times in Tn501 DNA.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>First base of recognition site</th>
</tr>
</thead>
<tbody>
<tr>
<td>AatII</td>
<td>1288</td>
</tr>
<tr>
<td>AccI</td>
<td>1885</td>
</tr>
<tr>
<td>AflIII</td>
<td>4763</td>
</tr>
<tr>
<td>AflIII</td>
<td>158</td>
</tr>
<tr>
<td>AsuII</td>
<td>603</td>
</tr>
<tr>
<td>Aval</td>
<td>695</td>
</tr>
<tr>
<td>AvalII</td>
<td>2064</td>
</tr>
<tr>
<td>BclI</td>
<td>6692</td>
</tr>
<tr>
<td>BstXI</td>
<td>284</td>
</tr>
<tr>
<td>DraII</td>
<td>2114</td>
</tr>
<tr>
<td>DraIII</td>
<td>4231</td>
</tr>
<tr>
<td>EcoRI</td>
<td>13</td>
</tr>
<tr>
<td>EspI</td>
<td>33</td>
</tr>
<tr>
<td>GdiI</td>
<td>6705</td>
</tr>
<tr>
<td>HgiEII</td>
<td>850</td>
</tr>
<tr>
<td>HgiHI</td>
<td>2064</td>
</tr>
<tr>
<td>HindIII</td>
<td>136</td>
</tr>
<tr>
<td>MaeI</td>
<td>68</td>
</tr>
<tr>
<td>NarI</td>
<td>1301</td>
</tr>
<tr>
<td>NdeI</td>
<td>2341</td>
</tr>
<tr>
<td>NheI</td>
<td>1524</td>
</tr>
<tr>
<td>NotI</td>
<td>3273</td>
</tr>
<tr>
<td>NruI</td>
<td>4352</td>
</tr>
<tr>
<td>PvuII</td>
<td>4852</td>
</tr>
<tr>
<td>SalGI</td>
<td>1885</td>
</tr>
<tr>
<td>Sphi</td>
<td>2498</td>
</tr>
<tr>
<td>Stul</td>
<td>6705</td>
</tr>
<tr>
<td>XhoII</td>
<td>1319</td>
</tr>
</tbody>
</table>

The enzymes from (42) which do not cleave Tn501 are: AhaIII, ApeI, AvalIII, AvrII, BamHI, BglII, BstEII, CiaI, EcoRV, HpaI, KpnI, MluI, NcoI, PvuII, PstI, RvuI, RsI, SalI, SacI, SacII, SauI, Scal, SfiI, SmaI, SnaI, SpeI, SspI, ThhIII, XbaI, XhoI and XmnI. (Isoschizomers are not included in this table.)

that transposase, which specifically recognises the long-range inverted repeats, may not contain the same type of DNA binding motif found in proteins that recognise short-range symmetrical sequences. Recent studies (31) have shown that the binding of Tn3 transposase to the inverted repeats of the transposon is ATP-dependent. This further argues against the DNA binding site being homologous to those of the ATP-independent DNA binding proteins. There is no sequence with obvious homology to known ATP-binding sites (30).

Complementation studies (4,5) have shown that, although the cleavage of the transposon terminus is precise, some transposases can recognise the terminal inverted repeats of closely-related transposons. For example, Tn501 and Tnl721 TnpA− mutants can be complemented by a functional
The hierarchy of complementation between Tn501, Tn1721, Tn21 and related transposons, makes the Tn3-related transposases powerful systems for the study of DNA-protein interaction. The availability of the primary sequences of two transposases of this family, and the gene sequences from which they are derived, will help in the design of experiments to identify those regions of the Tn3-related transposases required for specificity and for catalysis.

The DNA sequence of Tn501 and expression of the transposition genes

The DNA sequence of Tn501 is now completely defined, and examination of the full sequence has revealed several features relevant to the detailed biology of Tn501. Some of these have been discussed previously (3,7,14,23,26,27). This discussion is limited to those features which may affect the expression of the transposition functions. Some of the sequences discussed below (between positions 4231 and 5398) were presented by Diver et al (23).

In Tn3 and some other Tn3-like elements, such as gamma-delta, the transposition genes, tnpR and tnpA, are divergently transcribed, and the transcription of both genes is repressed by the binding of the tnpR gene product at the res site lying between the genes (33). In Tn501 the transposition genes are in the order res-tnpR-tnpA, and they are transcribed from the same strand as, and distal to, the inducible mercury resistance genes. The same gene order occurs in the closely-related transposons Tn1721 and Tn21. In Tn1721 there are three sequences in the res site that will bind resolvase. These have been identified by footprinting the protein-DNA complexes (34) and they are highly conserved in Tn501 (positions 4603-4730; given as 373-502 in ref. 23). The res sites of Tn501 and Tn1721 can participate in hybrid, reciprocal recombination catalysed by the Tn501 tnpR gene product (35). Tn1721 contains a promoter immediately in front of the tnpR gene (34). This promoter is in a
position to be regulated by the tnpR gene product binding at the res site, and in Tn1721 there is a reduction of 30% in expression of a gene expressed from this promoter when tnpR is supplied in trans (36). The proposed -35 and -10 sequences are of this promoter are conserved in Tn501 (at positions 4686-4691 and 4708-4713, respectively; 458-463 and 480-485 in ref. 23). We therefore assume that binding of resolvase to DNA may regulate the transcription of the tnpR gene in Tn501.

There is no sequence readily identifiable as being homologous to the consensus sequences of E. coli promoters (37) between the -35 and -10 regions of the presumptive tnpR promoter and the start of the tnpA gene. In Tn1721 the tnpA gene promoter is very weak, and could only be detected in the transposase-mediated transposition reaction (36). The high degree of homology between Tn1721 and Tn501 in this region implies that the Tn501 tnpA promoter would also be very weak. However, Kitts et al (38) showed that transposition of Tn501 in the absence of mercuric salts occurred at a low frequency and that the products were cointegrates, whereas in the presence of mercuric salts transposition occurred at a higher frequency and the products were resolved. This was taken as evidence that the tnpA gene has its own promoter, but that the tnpR gene does not, and that transcription of the transposition genes is induced by mercury, presumably by read-through from the mer genes. It is difficult to reconcile the data of Kitts et al (38) with the presence of a sequence in Tn501 which has identical -35 and -10 sequences and which shows 15/17 identity in the spacer region to the known promoter in Tn1721. We have found no structural features in Tn501 which are not present in Tn1721 which may explain this, and we are investigating the transcription of the transposition genes in more detail.

ACKNOWLEDGEMENTS

We thank Dr H. Muirhead for making protein structure prediction programs available to us, Dr H.C. Watson for providing computer facilities, Drs M.J. Bishop and G.G. Kneale for their help in using the Cambridge Nucleotide Sequence Data Library Service, K. Weston for instructing JNW in the ever-evolving DNA sequencing methods, and Drs P.M. Bennett, S.E. Halford and P.A. Lund for helpful discussion. This work was supported by grants from the MRC to NLB, who is a Royal Society EPA Cephalosporin Fund Senior Research Fellow.
REFERENCES