Volume 14 Number 11 1986 Nucleic Acids Research

The *trfB* region of broad host range plasmid RK2: the nucleotide sequence reveals *incC* and key regulatory gene *trfB/korA/korD* as overlapping genes

Christopher M.Thomas* and Christopher A.Smith

Department of Genetics, University of Birmingham, PO Box 363, Birmingham B15 2TT, UK

Received 26 March 1986; Revised and Accepted 12 May 1986

ABSTRACT

We report the nucleotide sequence of the trfB region of broad host range plasmid RK2. This region encodes the following loci: trfB, identical to korA and korD, which encodes a key transcriptional repressor of certain RK2 operons; incC, which appears to be involved in plasmid maintenance, possibly through post-transcriptional regulation of trfA product levels; the start of korB, which encodes a second transcriptional repressor of operons involved in stable inheritance of RK2. These loci are expressed as part of the trfB operon. In combination with deletion analysis, transcriptional and translation fusions and 'maxicell' analysis of polypeptides, the DNA sequence allows a number of conclusions to be drawn. First, the korB ORF start codon overlaps the incC ORF stop codon, suggesting the possibility of translational coupling between these two genes. Second, the trfB ORF lies entirely within the first third of the incC ORF using a different phase. Third, the incC ORF appears to contain a second transcriptional start whose function appears to be coupled to translation of the trfB ORF. Analysis of codon usage in the region of overlap between incC and trfB suggests that the incC gene may have evolved before the trfB gene. Determination of the DNA sequence of a mutant in which the product of trfB is rendered defective for transcriptional repression reveals an amino acid alteration within a region of this polypeptide which exhibits homology to the αhelix-turn-αhelix motif characteristic of many DNA binding proteins, and which is probably responsible for recognition of the trfB operator by this protein.

INTRODUCTION

Plasmid RK2, which belongs to *Escherichia coli* incompatibility group P, is capable of efficient transfer between and stable maintenance in most Gram negative bacterial species (for a review see reference 1). Deletion analysis originally produced mini replicons containing three regions of RK2 (2). One region contains $oriV_{RK2}$, from which vegetative replication proceeds unidirectionally (3). A second region contains trfA, at least one of whose products is essential for replication from $oriV_{RK2}$ (4,2,5). Together, $oriV_{RK2}$ and trfA are sufficient for replication (5) and form a replicon with a broad host range (6,7,8). The third region in the original

© IRL Press Limited, Oxford, England.

mini replicons contains a locus originally designated trfB, which is identical to korA (9,24) and korD (9), whose product represses transcription of the trfA operon, encoding both trfA (9,10,11,12) and kilD (also designated kilB1) (9,13). The kilD locus is one of a number of RK2 loci which are either host lethal (kilA, B, C) or which interfere with plasmid maintenance (kilD) if unregulated by the respective kor (kil-override) function (14,9,13). Since the trfA region in the original mini plasmids contains kilD as well as trfA, the trfB region was necessary to suppress the KilD$^+$ phenotype rather than being essential for replication. When the trfB region is deleted, mutants which are phenotypically KilD$^-$ can be isolated and the one mutant of this sort which we have analysed further shows a point mutation which reduces the unregulated strength of the promoter for the trfA operon (5, 15).

The trfB locus is part of an operon, the trfB operon, which also codes for both incC (defined below) and korB (16). Transcription of this operon is autogenously repressed by the products of both trfB and korB (17,18) both of which are rather general regulatory elements. Thus, in addition to its autorepression and its repression of trfA operon transcription (korD, 12) the trfB product represses transcription of kilA (korA, 18) and appears to be involved in regulation of kilC (19). Similarly the korB product is also the prime transcriptional repressor of the kilB operon (12,15), is a second repressor for the trfA operon (12,15) and probably also represses kilA (15,18). While kil and kor genes are not essential for replication, the coordination of their control with that of the replication gene trfA suggests a functional relationship. For this reason, and because of their similarity to auxiliary stable inheritance functions found on other plasmids, we have suggested that the kil and kor genes may be involved in aspects of plasmid inheritance, such as partitioning to daughter cells and coupling of host cell division to plasmid multiplication (20). Thus the products of at least two genes of the trfB operon play a central role in regulating expression of an important set of coordinately controlled plasmid genes.

The third known gene in the trfB operon is incC (21,16) whose product can interfere in trans with the maintenance of parental RK2 plasmids but not with plasmids lacking korB (21,22). A clue to the action of incC comes from studies on a mutant plasmid, pRK2501ts3, derived from a mini RK2 plasmid lacking korB (5,23). The mutation reduces the transcriptional repressor activity of the trfB product resulting in partial derepression of

both the trfB and trfA operons. The plasmid is unstable at 30°C and lost rapidly at 42°C. This instability can be suppressed by deletions inactivating incC, implicating derepression of incC expression in the plasmid maintenance defect at 42°C. At 42°C in this plasmid the trfA protein concentration is reduced, an effect reversed by deletions in incC (23). How incC derepression has this effect, if there is a direct connection, is not clear at present, but it does seem likely that the incC product represents an additional level of control for plasmid replication and maintenance.

In addition to the question of the role of incC it is intriguing that our previous analysis of the trfB region (16) suggested that the incC gene might overlap the transcriptional regulator gene trfB, a suggestion that was compatible with the published DNA sequence of the trfB gene (24) since the putative korA open reading frame (ORF) was found to be encompassed within an ORF which extended beyond the published sequence. As part of our study of the stable inheritance of IncP plasmids we have determined the DNA sequence of the minimal trfB region found in the original mini replicons. This consists of two HaeII fragments previously designated B_1 and B_2 (2). The sequence has allowed us to determine in detail the organization of the trfB, incC and korB genes and to study some aspects of their expression.

MATERIALS AND METHODS

Bacterial strains and plasmids

Escherichia coli K12 strains C600K$^-$ (thr-1 leu-6 thi-1 lacY1 supE44 tonA21 galK) from M. Rosenberg and MV10 (thr-1 leu-6 thi-1 lacY1 supE44 tonA21 trpE5) from D.R. Helinski were used as standard hosts while CSR603 (25), a recA1 phr-1 derivative of AB1886 (thr-1 leu-6 thi-1 lacY1 galK2 ara-14 xyl-5 mtl-1 proA2 his-4 argE3 str-31 tsx-33 sup-37 uvrA6) from D.R. Helinski was used for producing 'maxicells' (see below).

Plasmids not previously published are either described in the legend to Fig. 1 or are shown in Fig. 3. Plasmids described previously are referenced in the text.

Growth media

L-broth (26) was used as standard liquid medium. 1.5% (w/v) agar was added for solid medium. M9 casamino acids medium (M9.CAA; 26) was used to provide a tryptophan-free medium. Bacteria were grown at 37°C. Antibiotic resistance was selected by the addition of penicillin-G for Pn^r (150

μg/ml in liquid, 300 μg/ml in solid medium), kanamycin sulphate for Km^r (50 μg/ml) and tetracycline hydrochloride for Tc^r (10 μg/ml).

Plasmid DNA isolation, analysis and manipulation

For rapid analysis of plasmid DNA the alkaline-SDS method of Birnboim and Doly (27) was used with slight modifications (9). Purification on CsCl/Ethidium bromide gradients was as described previously (9). Agarose gel electrophoresis was carried out by standard techniques. Restriction and DNA modifying enzymes were purchased from Bethesda Research Laboratories, nbl enzymes or BCL and used essentially according to the manufacturer's instructions.

Bal31 deletions were generated as follows. Approximately 30 μg of linearized plasmid DNA was dissolved in 1 ml of Bal31 buffer with Bovine serum albumin at 250 μg/ml (28) and after addition of 2 units of Bal31 enzyme, aliquots were withdrawn after 0.5, 1, 2, 4 and 6 min. incubation at 30°C, phenol extracted twice to stop the reaction, ether extracted and then the DNA precipitated. After analysis of a sample from each time point to estimate the extent of digestion, chosen DNA samples were 'filled in' with DNA polymerase I large fragment and dNTPs and then ligated to unphosphorylated 10 bp HindIII linkers according to the method of Lathe et al. (29) to insert a single linker in the gap between the ends generated after deletion.

Transposon insertion mutagenesis

Transposon insertion mutagenesis with Tn1725 (30) was carried as previously described (10).

DNA sequence determination

DNA sequences were determined by the method of Maxam and Gilbert (31) with minor modifications as described previously (11). The strategy for sequencing the trfB region is described in Fig. 1.

'Maxicell' analysis of plasmid encoded polypeptides.

This was carried out by standard procedures (25) with slight modifications as described previously (10).

RESULTS AND DISCUSSION

Nucleotide sequence of the trfB region

The strategy used for determining the DNA sequence of both HaeII B_1 and B_2 which constitute the trfB region is shown in Fig. 1. It relied largely on the use of insertion derivatives which place convenient restriction sites adjacent to a variety of internal segments. The only

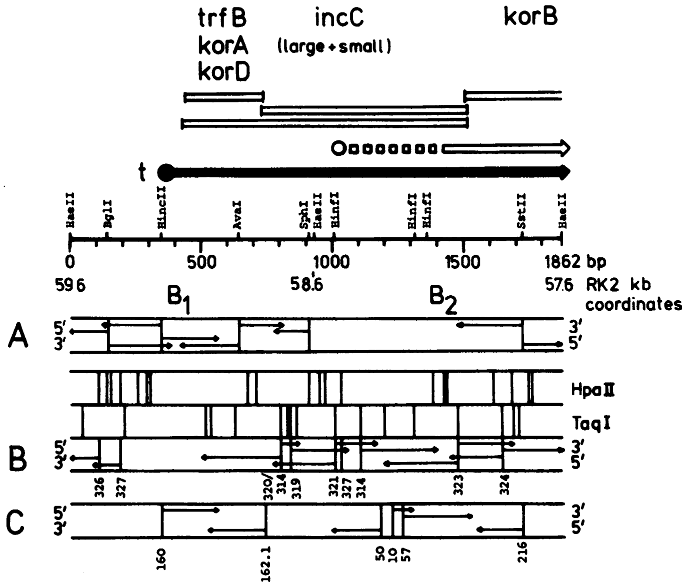

FIG. 1. Physical and genetic map of the trfB region of RK2 showing the strategy for determining the DNA sequence. The region shown runs from kilobase (kb) coordinates 59.6 to 57.6 (defined clockwise as conventionally drawn from the unique EcoRI on the 60 kb genome) or 54.0 to 56.0 when the previous 0-56.4 kb coordinates are used. However, our analysis shows this region to consist of only 1862 bp rather than 2000 bp as previously estimated. Genetic organization based on this and previous work is shown as: ●, known promoter (15); o, putative promoter (15); [arrow], mRNA transcript (15,24); t, putative transcriptional terminator; [bar], open reading frames (ORF) known be translated (15; this work). DNA sequence was obtained in the following way. First, by 5′ and/or 3′ labelling (5′ ends were labelled using T4 polynucleotide kinase and $[\gamma-^{32}P]$-ATP; 3′ ends were labelled with terminal transferase and $[\alpha-^{32}P]$-ddATP) after digestion of DNA with restriction endonucleases which cut just once in this region (A). Second, by 5′ labelling of EcoRI or HindIII fragments obtained from plasmids pCT319.1*, pCT320, pCT321*, pCT323*, pCT324*, pCT326, pCT327. Plasmids marked * were previously described (9,16); the rest are analogous. In each, pBR322, cleaved at the ClaI site (which lies between the EcoRI and HindIII sites) is inserted at one or other TaqI or HpaII site in the trfB region as shown (B). Third, by 5′ labelling of appropriate sites in plasmids pCAS160 (15), pCAS162.1 (a deletion derivative of pCAS162 bringing an internal segment of trfB close to the pBR322 EcoRI site; pCAS162 itself being an XhoI-SalI deletion of pCT321, ref. 9,16) and pCAS216 (the EcoRI to SstII trfB region fragment from pCT16 (2) inserted between EcoRI and ClaI sites of pBR322 with the aid of an SstII to TaqI 'linker' segment of DNA from R751) with subcloned segments of the trfB region (C). Fourth, by 5′ labelling at the EcoRI sites introduced by insertion of Tn1725 (30) into pCT143 (Fig. 3) at the locations shown here and in Fig. 2 to give pCT143::Tn1725 #50, #10 and #57, (C). Sequence for both strands was obtained except for a short region which agrees with the sequence of korA published previously (24).

```
HaeII
GGCGCTGGGTTTTACGCCAAGGGGAGAAGCGGCATACGCCATCGAGCGGTCAGCCTATGGTGAGCCGAGGTCCAGCATTGCGAAGCAGTACCAGCAGGAA
CCGCGACCCAAAATGCGGTTCCCCTCTTCGCCGTATGCGGTAGCTCGCCAGTCGGATACCACTCGGCTCCAGGTCGTAACGCTTCGTCATGGTCGTCCTT
                                                                                                      100

                                          BglI
TGGAACCGGAAGGCGGCGACCTGGTGGGCGATGGCCGGTGTGGCCGGCATCATCGGCGCGATCCTGGCGGCGGCGGCAACCGGCTTTGTTGGGCTGGCAG
ACCTTGGCCTTCCGCCGCTGGACCACCCGCTACCGGCCACACCGGCCGTAGTAGCCGCGCTAGGACCGCCGCCGCCGTTGGCCGAAACAACCCGACCGTC
     326                                                                       327                    200

                                                                                    <------t?---
                                                                                         IRIII
TGTCGATCCGCAACCGAGTGAAGCGCGTGCGCGACCTGTTGGTGATGGAGCCGGGTGCAGAGCCATAAGCGGCAAGAGACGAAGCCCGGTTTCCGGGCTT
ACAGCTAGGCGTTGGCTCACTTCGCGCACGCGCTGGACAACCACTACCTCGGCCCACGTCTCGGTATTCGCCGTTCTCTGCTTCGGGCCAAAGGCCCGAA
                                                                                                      300

------>                                     -35                   -10
                              IRI          HincII                 IRII
TTGTTTTGTTACGCCAAGGACGAGTTTTAGCGGCTAAAGGTGTTGACGTGCGAGAAATGTTTAGCTAAACTTCTCTCATGTGCTGGCGGCTGTCACCGCT
AACAAAACAATGCGGTTCCTGCTCAAAATCGCCGATTTCCACAACTGCACGCTCTTTACAAATCGATTTGAAGAGAGTACACGACCGCCGACAGTGGCGA
                                                                                                      400

                                P364(incC) P101(trfB)
                                M  G  V  I  H  E  E  T  A  Y  R  K  P  V  P  G  G  D  P  G  A  G  S  G
                     SD           SD          M  K  K  R  L  T  E  S  Q  F  Q  E  A  I  Q  G  L  E  V  G
ATGTTCAACCAAGGCGCGGAGCAAATTATGGGTGTTATCCATGAAGAAACGGCTTACCGAAAGCCAGTTCCAGGAGGCGATCCAGGGGCTGGAAGTGGGG
TACAAGTTGGTTCCGCGCCTCGTTTAATACCCACAATAGGTACTTCTTTGCCGAATGGCTTTCGGTCAAGGTCCTCCGCTAGGTCCCCGACCTTCACCCC
                                                                                                      500

A  A  D  H  R  D  S  A  G  R  L  S  R  H  E  A  T  G  D  V  R  N  V  A  G  T  D  Q  G  R  S  V  A  S
 Q  Q  T  I  E  I  A  R  G  V  L  V  D  G  K  P  Q  A  T  F  A  T  S  L  G  L  T  R  G  A  V  S  Q
CAGCAGACCATCGAGATAGCGCGGGGCGTCTTAGTCGATGGGAAGCCACAGGCGACGTTCGCAACGTCGCTGGGACTGACCAGGGGCGCAGTGTCGCAAG
GTCGTCTGGTAGCTCTATCGCGCCCCGCAGAATCAGCTACCCTTCGGTGTCCGCTGCAAGCGTTGCAGCGACCCTGACTGGTCCCCGCGTCACAGCGTTC
                                                                                                      600

                                          AvaI
  G  A  S  R  V  G  R  V  R  G  Q  E  L  A  R  G  V  R  A  G  N  G  G  S  A  G  T  S  G  V  H  R  P
A  V  H  R  V  W  A  A  F  E  D  K  N  L  P  E  G  Y  A  R  V  T  A  V  L  P  E  H  Q  A  Y  I  V  R
CGGTGCATCGCGTGTGGGCCGCGTTCGAGGACAAGAACTTGCCCGAGGGGTACGCGCGGGTAACGGCGGTTCTGCCGGAACATCAGGCGTACATCGTCCG
GCCACGTAGCGCACACCCGGCGCAAGCTCCTGTTCTTGAACGGGCTCCCCATGCGCGCCCATTGCCGCCAAGACGGCCTTGTAGTCCGCATGTAGCAGGC
                                                                                                      700

                            SD         P259
E  V  G  S  G  R  Q  E  K  T  G  N  Q  T  M  K  T  L  V  T  A  N  Q  K  G  G  V  G  K  T  S  T  L
 K  W  E  A  D  A  K  K  K  Q  E  T  K  R  D
GAAGTGGGAAGCGGACGCCAAGAAAAAACAGGAAACCAAACGATGAAAACTTTGGTCACGGCCAACCAGAAAGGCGGCGTCGGCAAGACTTCGACCCTTG
CTTCACCCTTCGCCTGCGGTTCTTTTTTGTCCTTTGGTTTGCTACTTTTGAAACCAGTGCCGGTTGGTCTTTCCGCCGCAGCCGTTCTGAAGCTGGGAAC
                                            <-|                                           320/314     800
                                            162.1

V  H  L  A  F  D  F  F  E  R  G  L  R  V  A  V  I  D  L  D  P  Q  G  N  A  S  Y  T  L  K  D  F  A  T
TGCATCTTGCCTTCGACTTTTTCGAGCGCGGCTTGCGGGTTGCCGTGATCGACCTGGACCCCCAGGGCAATGCGTCCTACACGCTCAAGGACTTTGCTAC
ACGTAGAACGGAAGCTGAAAAAGCTCGCGCCGAACGCCCAACGGCACTAGCTGGACCTGGGGGTCCCGTTACGCAGGATGTGCGAGTTCCTGAAACGATG
                      319                                                                             900

       SphI              HaeII
  G  L  H  A  S  K  L  F  G  A  V  P  A  G  G  W  T  E  T  A  P  A  A  G  D  G  Q  A  A  K  L  A  L
CGGCCTGCATGCAAGCAAGCTGTTCGGCGCTGTCCCTGCCGGCGGCTGGACCGAAACCGCACCCGCAGCCGGCGACGGCCAGGCCGCGCGCCTCGCCCTC
GCCGGACGTACGTTCGTTCGACAAGCCGCGACAGGGACGGCCGCCGACCTGGCTTTGGCGTGGGCGTCGGCCGCTGCCGGTCCGGCGCGCGGAGCGGGAG
                                                                                                      1000

  HinfI
  I  E  S  N  P  V  L  A  N  A  E  R  L  S  L  D  D  A  R  E  L  F  G  A  N  I  K  A  L  A  N  Q  G
ATCGAGTCCAACCCGGTACTGGCGAACGCCGAACGGCTGTCGCTGGACGACGCCCGCGAGCTGTTCGGGGCGAACATCAAGGCCCTGGCGAACCAAGGCT
TAGCTCAGGTTGGGCCATGACCGCTTGCGGCTTGCCGACAGCGACCTGCTGCGGGCGCTCGACAAGCCCCGCTTGTAGTTCCGGGACCGCTTGGTTCCGA
 321        327                                                                                       1100

F  D  V  C  L  I  D  T  A  P  T  L  G  V  G  L  A  A  A  L  F  A  A  D  Y  V  L  S  P  I  E  L  E  A
TCGACGTGTGCCTGATCGACACGGCCCCGACCCTTGGCGTCGGCCTGGCGGCCGCCCTCTTCGCGGCCGACTATGTGCTGTCCCCCATCGAGCTTGAGGC
AGCTGCACACGGACTAGCTGTGCCGGGGCTGGGAACCGCAGCCGGACCGCCGGCGGGAGAAGCGCCGGCTGATACACGACAGGGGGTAGCTCGAACTCCG
314                                                                         50                        1200
```

```
   Y  S  I  Q  G  I  K  K  M  V  T  T  I  A  N  V  R  Q  K  N  A  K  L  Q  F  L  G  M  V  P  S  K  V
GTACAGCATCCAGGGCATCAAGAAGATGGTCACGACCATTGCGAACGTGCGCCAGAAGAACGCCAAGCTGCAATTCCTTGGCATGGTGCCCAGCAAGGTC
CATGTCGTAGGTCCCGTAGTTCTTCTACCAGTGCTGGTAACGCTTGCACGCGGTCTTCTTGCGGTTCGACGTTAAGGAACCGTACCACGGGTCGTTCCAG
         .      10 .         .         .         .    57   .         .         .         .         .   1300

      HinfI                                           HinfI
  D  A  R  N  P  R  H  A  R  H  Q  A  E  L  L  A  A  Y  P  K  M  M  I  P  A  T  V  G  L  R  S  S  I
GATGCGCGGAATCCGCGCCACGCGCGCCACCAAGCCGAGCTGCTGGCCGCGTACCCCAAGATGATGATTCCGGCCACCGTTGGCCTGCGCAGCAGCATCG
CTACGCGCCTTAGGCGCGGTGCGCGCGGTGGTTCGGCTCGACGACCGGCGCATGGGGTTCTACTACTAAGGCCGGTGGCAACCGGACGCGTCGTCGTAGC
         .         .         .         .         .         .         .         .         .         .   1400

A  D  A  L  A  S  G  V  P  V  W  K  I  K  K  T  A  A  R  K  A  S  K  E  V  R  A  L  A  D  Y  V  F  T
CCGATGCCCTCGCATCCGGTGTGCCGGTCTGGAAGATCAAGAAAACGGCCGCGCGCAAGGCATCGAAAGAGGTTCGCGCCCTGGCTGATTACGTGTTCAC
GGCTACGGGAGCGTAGGCCACACGGCCAGACCTTCTAGTTCTTTTGCCGGCGCGCGTTCCGTAGCTTTCTCCAAGCGCGGGACCGACTAATGCACAAGTG
         .         .         .         .         .         .  323    .         .         .         .   1500

                       P119'(korB')
       SD              |--->
                       M  T  A  A  Q  A  K  T  T  K  K  N  T  A  A  A  A  Q  E  A  A  G  A  A  Q  P  S
  K  M  E  I  S  Q  D
GAAGATGGAGATTTCCCAATGACTGCGGCTCAAGCCAAGACCACCAAGAAAAACACCGCTGCGGCCGCTCAGGAAGCCGCAGGCGCGGCGCAGCCGTCCG
CTTCTACCTCTAAAGGGTTACTGACGCCGAGTTCGGTTCTGGTGGTTCTTTTTGTGGCGACGCCGGCGAGTCCTTCGGCGTCCGCGCCGCGTCGGCAGGC
         .         .         .         .         .         .         .         .         .         .   1600

G  L  G  L  D  S  I  G  D  L  S  S  L  L  D  A  P  A  A  S  Q  G  G  S  G  P  I  E  L  D  L  D  L  I
GCCTGGGGTTGGATAGCATCGGCGACCTGTCGAGCCTCCTGGACGCTCCTGCGGCGTCTCAGGGCGGTTCCGGCCCTATCGAGCTGGACCTGGACCTGAT
CGGACCCCAACCTATCGTAGCCGCTGGACAGCTCGGAGGACCTGCGAGGACGCCGCAGAGTCCCGCCAAGGCCGGGATAGCTCGACCTGGACCTGGACTA
         .         .         .         .         .         .         .         .         .         .   1700
                              324

  D  E  D  P  M  C  P  R  T  A  D  N  P  G  F  S  P  E  S  I  A  E  I  G  A  T  I  K  E  R  G  V  K
                    SstII
CGACGAAGATCCGCATCAGCCGCGGACGGCCGACAACCCCGGCTTTTCCCCGGAGAGCATCGCGGAAATCGGTGCCACGATCAAAGAGCGCGGGGTGAAG
GCTGCTTCTAGGCGTAGTCGGCGCCTGCCGGCTGTTGGGGCCGAAAAGGGGCCTCTCGTAGCGCCTTTAGCCACGGTGCTAGTTTCTCGCGCCCCACTTC
         .         .         .         .         .         .         .         .         .         .   1800

 S  P  I  S  V  R  E  N  C  E  C  P  G  R  Y  I  I  N  H  G
                                                        HaeII
TCACCCATTTCGGTGCGCGAGAACCAGGAGCAGCCGGGCCGCTATATCATCAATCACGGCGC
AGTGGGTAAAGCCACGCGCTCTTGGTCCTCGTCGGCCCGGCGATATAGTAGTTAGTGCCGCG
         .         .         .         .         .         .         .         .         .         .   1900
```

FIG. 2. Nucleotide sequence of the trfB region and amino acid sequences of predicted polypeptides. The target sites duplicated by Tn1725 insertions (boxes containing 5bp) used for sequence determination are shown as well as the TaqI and HpaII restriction sites (boxes containing 4bp) where pBR322 is inserted in the plasmids shown under (B) FIG. 1. Promoters (-10 and -35 regions); inverted repeats representing operators (IR I and II) and a putative terminator (IR III); and putative Shine Dalgarno sequences (SD) are shown. The standard single letter code for amino acids is used to show the polypeptide sequences.

region for which both strands were not determined was part of the coding region for trfB, since our data agreed with the sequence published for this region (24). We found an additional base (14), not reported by Bechhofer and Figurski (24), in a region for which both strands were determined. The G+C content of the region sequenced is 63% and may explain the presence of a number of sequences which exhibit 'compression' on DNA sequencing gels due to the formation of G+C-rich intrastrand stem-loop structures which we found considerable difficulty in disrupting. One such sequence runs from position 1313-1319, but we were eventually successful in identifying

distinct bands for every base on a 20% acrylamide gel with sequencing reactions carried out on a DNA fragment whose kinase labelled 5′ end was about 60 bp away from the region of potential compression. The nucleotide sequence is shown in Fig. 2 along with the open reading frames corresponding to trfB , incC and korB as demonstrated below.

We have previously identified a polypeptide of M_r = 13-14,000 as the product of the trfB locus (15). Given the potential for SDS- PAGE to give poor estimates for molecular weight it seems likely that this represents the product of the 101 codon ORF identified by Bechhofer and Figurski (24) which would give a polypeptide of 11,306 daltons. This open reading frame starts at base 441 and ends with the codon beginning at base 744.

The trfB region is also known to code for the N-terminal section of korB which is transcribed as part of the same operon as trfB and incC (16). The region coding for korB appears to span the SstII site (9). Analysis indicates that only one ORF of sufficient length extends across both the SstII site and the HaeII site at the end of fragment HaeII B_2. This begins with a codon starting at base 1519. This position agrees with our previous estimates for the coding region for korB (16). The ATG at the start of this ORF is preceded by a potential Shine Dalgarno (SD) sequence (GGAG) separated from the ATG by 8 bases. This ATG overlaps the putative stop codon for the incC ORF (see below).

The major incC product is a polypeptide of M_r = 38,000 (16) whose C-terminal is encoded somewhere between positions 1465 and 1630. This is indicated by the production of truncated polypeptides from a plasmid with the trfB region disrupted at position 1465, as in pCT323, but not from a plasmid with disruption at position 1630, as in pCT324 (16). There is only one ORF which is long enough to code for the M_r = 38,000 polypeptide. Of the three possible start codons (GTG, position 380-382; ATG, position 401-403 and ATG, position 428-430) only the ATG at position 428-430 is preceded by a good S.D. sequence (GGAG, 6 bases upstream of ATG). This ORF extends to a stop codon at position 1520-1522, which is consistent with our previous conclusion that the incC C-terminal is encoded between positions 1465 and 1630. This ORF is the same as that noted by Bechhofer and Figurski (24), proceeding through their korA sequence (which extends as far as position 836 in the sequence reported here). This ORF would give a 364 amino acid polypeptide of MW = 38,134 daltons which agreed with the SDS-PAGE estimate.

The incC ORF contains two translational starts

The analysis of truncation products referred to above (16) indicates that there are additional incC associated polypeptides of which the most prominent is one of M_r = 25,000 with a less intense band or bands at M_r = 30,000. These can be seen in Fig. 3, track A1 which shows polypeptides produced by pCT143 (23), which consists of the wild type trfB region joined to pBR322, analysed in maxicells. Fig. 3, track A2 shows that a plasmid, pCT640, isogenic with pCT143 except for a spontaneous deletion (from pRK2501ts3R_0; 5,23) of 70 bp with its end points in the directly repeated sequence CGCG at positions 985-8 and 1055-8, no longer produces the M_r = 38,000, 30,000 and 25,000 polypeptides. From the sequence of this deletion it is predicted that the longest truncated polypeptide which the incC ORF would give would be of MW = 20,973, in agreement with the major truncation product observed in Fig. 3, track A2.

The secondary incC polypeptides could result either from a secondary translation initiation signal within the incC ORF or from processing of the primary 38,134 dalton polypeptide. Analysis of the incC ORF reveals the presence of an ATG codon beginning at position 743 and preceded by a putative Shine Dalgarno sequence (GGA separated by 9 bases from the ATG) which if used would result in a polypeptide of MW = 27,533. Interestingly, this ATG overlaps the stop codon for the trfB ORF in one of the standard ways (ATGA) observed for translationally coupled ORFs (32). As a result of the experiments described below we believe that this second translational start results in the M_r = 25,000 polypeptide. No other likely translational starts, which could give rise to the M_r = 30,000 polypeptide(s), were found, suggesting that these are likely to be the products of processing the M_r = 38,000 polypeptide.

If the smaller incC polypeptides are proteolytic products of a single initial translation product then inactivation of the start site at position 428-430 should result in loss of all incC polypeptides while if any are primary translation products they should still be produced under these circumstances. This was investigated as described below.

Plasmid pCT143 (Fig. 4) is a pBR322 derivative carrying the wild type trfB region and gives all the known trfB region polypeptide products (Fig. 3, line A1). A derivative (#12) of pCT143 carrying transposon Tn1725, whose insertion generated direct repeats of the pentamer TATGT at position 400-404 which flank Tn1725 in this derivative, was used to create pCT650, in which the trfB promoter is deleted (Fig. 4). Plasmid pCT650 does not

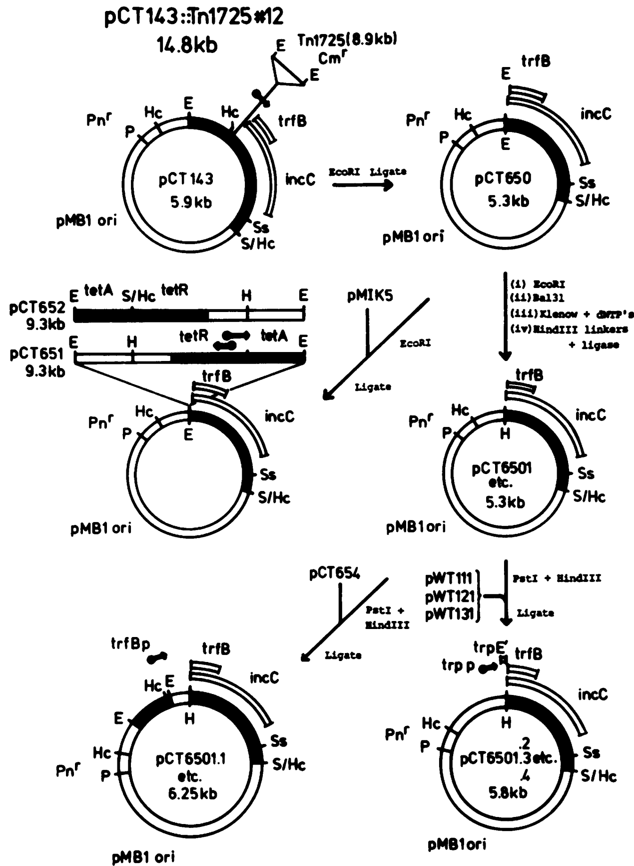

FIG. 3. Structure of pCT143 and its derivatives. Plasmid pCT143 consists of the SstII to EcoRI segment of RK2 (coordinates 57.8 to 60.0 kb) containing the trfB region; a 241 bp SstII to SalI fragment from the tetA gene of RK2 (as a 'linker') and the SalI to EcoRI section of pBR322 containing the pMB1 replicon and the Pn^r determinant. Plasmid pCT143:: Tn1725 #12 has Tn1725 (30) inserted as shown, the insertion having generated direct repeats of bases 400-404 as shown in Fig. 2. Plasmid pCT650 consists of the EcoRI fragment of pCT143::Tn1725 #12 carrying Pn^r, pMB1 ori and the trfB and incC ORFs but not the trfB promoter. Plasmids pCT651 and pCT652 were constructed by inserting into pCT650 an EcoRI fragment from pMIK5 (a Tn1723 insertion derivative of pCT87 (5)) in which tetA transcription, which is inducible by tetracycline due to the presence of the tetR gene, will activate adjacent genes. Plasmids pCT650.1 etc. were created by Bal31 deletion and insertion of a HindIII linker. '.1' derivatives were generated by insertion of a PstI to HindIII fragment from pCT654 (18) containing the wild type trfB promoter. '.2', '.3' and '.4' derivatives were generated by insertion of the PstI to HindIII trpE promoter fragment from pWT111, pWT121 and pWT131 respectively (33). Restriction sites are represented as follows: EcoRI, E; HincII, Hc; HindIII, H; PstI, P; SalI, S; SstII, Ss.

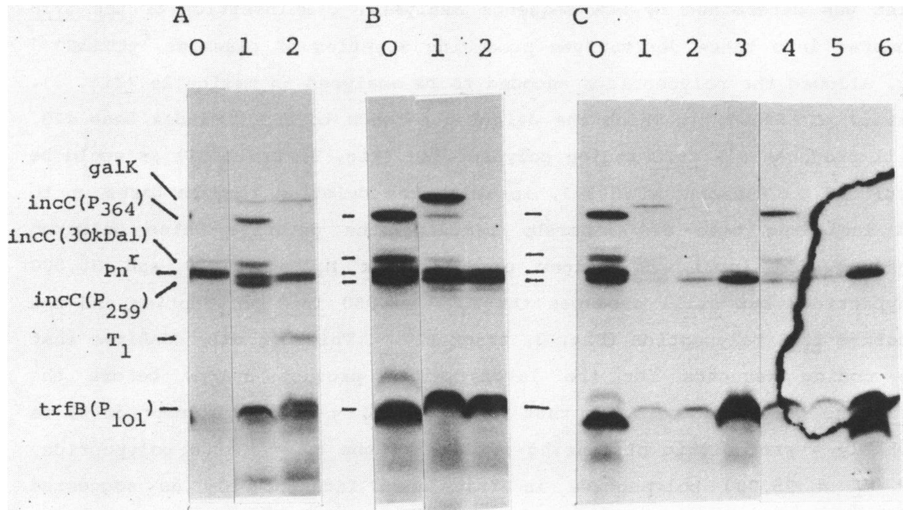

FIG. 4. Maxicell analysis of polypeptides. A. Plasmid pCT673 (23) is a galK promoter probe plasmid containing the pUC18 polylinker (38) upstream separated from galK by λt_{R1}. The EcoRI to SalI trfB segment of pCT143 (Fig. 3) was inserted into pCT673 to give pCT676 (23). Comparison of the polypeptides produced by pCT673 (track 0) and pCT676 (track 1) shows those produced by the trfB region. The incC-associated polypeptides are identified by their disappearance in track 2 which shows polypeptides produced by plasmid pCT677, isogenic with pCT676 except for a 70 bp deletion in incC (see text) which produces the truncation products shown (T_1, T_2, T_3). B. Loss of the large incC polypeptide as a result of a deletion removing the incC Shine Dalgarno sequence. Track 0 shows the polypeptides of pCT324 which serve to identify the trfB region polypeptides. Track 1 shows polypeptides produced by pCT6516.1 in which the incC SD is intact; track 2 shows those of pCT6503.1 in which the incC SD is deleted. C. Translational fusions of incC and trfB to trpE expression signals. Plasmid pCT324 shows the trfB region polypeptides (track 0). Polypeptides produced by pCT6516.2, pCT6516.3 and pCT6516.4 and pCT6509.2, pCT6509.3 and pCT6509.4 (see Fig. 3) are shown in tracks 1 to 6 respectively.

express an $IncC^+$, $KorA^+$ or $KorD^+$ phenotype. However, when a fragment carrying a promoter (Fig. 4) was inserted at the EcoRI site of pCT650, both the $IncC^+$ and $KorA^+$ phenotypes were obtained with the orientation which would give transcription across the trfB region, pCT651, but not with the opposite orientation, pCT652, (data not shown) indicating that the incC and trfB coding sequences are still intact in pCT650.

A series of deletions from the EcoRI site were generated with Bal31 exonuclease and a HindIII linker introduced at the end of each deletion (Fig. 4) producing a series of plasmids, pCT6501 etc. The deletion end

point was determined by DNA sequence analysis. Reinsertion of the trfB promoter into these derivatives producing a series of plasmids, pCT6501.1 etc. allowed the polypeptides encoded to be analysed in maxicells (Fig. 3). Plasmid pCT6516.1, in which the deletion extends to and includes base 410, still produces all trfB region polypeptides (Fig. 3, track B1) as would be predicted. Plasmid pCT6503.1, in which the deletion removes bases up to and including base 421 (thereby removing the putative Shine Dalgarno sequence for incC), no longer produces the M_r = 38,000 and 30,000 polypeptides but still produces the M_r = 25,000 incC polypeptide and the putative trfB polypeptide (Fig. 3, track B2). This not only confirms that the coding sequence for the largest incC product starts before the trfB gene but also indicates that while the M_r = 30,000 polypeptide(s) is probably a proteolytic processing product of the M_r = 38,000 polypeptide, the M_r = 25,000 polypeptide is translated independently as suggested above.

Translational coupling between trfB and the secondary start site within incC

While the M_r = 25,000 polypeptide appears to be translated independently of the larger incC polypeptides, the overlap of its start codon with the stop codon of trfB suggests that its translation may be coupled to that of trfB. To determine whether or not this is the case we created translational fusions in all three phases to boost translation of either large incC or trfB or neither and determined their effect on the expression of small incC. This was done using pCT6516 and pCT6509, two of the Bal31-generated deletion derivatives of pCT650. DNA of these plasmids was cut with PstI and HindIII and a PstI to HindIII fragment from either pWT111 or pWT121 or pWT131 (33) inserted, creating translational fusions in all three reading frames to the E.coli trpE ORF transcribed from the E.coli trp promoter. Fig. 3c shows the polypeptide products of the derivatives of pCT6516 (see above) and pCT6509, in which bases up to and including 443 are deleted. The results show that the translational fusions boost expression of either large incC ('.2' derivatives) or trfB ('.4' derivatives) as predicted from the DNA sequence, thus confirming the identification of ORFs. Small incC polypeptide is only boosted by fusions in frame with trfB. The large incC polypeptide is increased in size by the translational fusion in pCT6516.2 (compare tracks C0 and C1) as expected due to the N-terminal addition of 16 amino acids from the start of trpE and the trpE-incC junction. It is interesting that

even though the normal incC translation initiation signals should still exist in this plasmid their activity seems to be completely suppressed by the trpE translational start. A similar situation appears to exist with the trfB gene product where the appropriate derivative of pCT6509 produces a smaller fusion product than the pCT6516 derivative. In the pCT6509.2 the fusion results in a net gain of only 5 N-terminal amino acids to large incC since some are lost as a result of the Bal31 deletion. It therefore appears to be of approximately normal size. The M_r = 30,000 incC-associated polypeptide(s) is boosted only when the largest incC polypeptide is produced in large quantities. Its size however is not affected by the translational fusion suggesting that proteolysis, the likely source of this protein, removes the N-terminal end. That the small incC polypeptide (M_r = 25,000) is boosted only in conjunction with fusions in phase with trfB supports the suggestion that there is translational coupling between the expression of these two polypeptides. However, the biological significance of this coupling remains unclear.

Properties of trfB region-encoded polypeptides

The primary structure of the trfB polypeptide has been described previously (24). However, a fact not described previously is that this polypeptide shows a region (amino acids 35 to 56) which shows considerable homology in primary and potential secondary structure with the αhelix-turn-αhelix motif found in many regulatory DNA binding proteins (for a review see ref. 34). Not only does this region of the trfB polypeptide (TrfB) show ala, gly and val as expected at the highly conserved positions 5, 9 and 15 respectively but it also shows a distribution of hydrophobic and hydrophilic amino acids similar to that in λ repressor and Cro proteins which is consistent with αhelices with one side exposed for DNA binding and the other side facing in towards the centre of the protein. Further support for the importance of this region in DNA binding comes from our previously unpublished nucleotide sequence analysis on the trfB region of a mutant plasmid carrying a mutation (trfBts3; 5,23), which renders the trfB polypeptide at least partially defective with respect to transcriptional regulation at both 30°C and 42°C (23). The mutation results in a val→met alteration at the conserved position 15 in this region (amino acid number 51 in TrfB). This is due to a G→A transition at position 591, as determined by DNA sequence analysis after labelling wild-type and mutant DNA at the AvaI site in the trfB gene. In λ repressor and Cro the amino acids at positions 5 and 15 in this region are

in van der Waals contact and their interaction appears to be important for the precise αhelix-turn-αhelix structure. Replacement of valine by methionine at this position would alter the exact structure which could result in decreased binding strength of TrfBts3 for the operator DNA, explaining its poorer repressor activity.

The incC ORF shows a codon usage very similar to that observed for the genes of the trfA region of RK2 (11). This distribution shows a bias for codons ending in G or C. The G+C content of the third position within the codons is 68% compared with an overall G+C content of 63% for the region shown. The amino acid composition of the incC polypeptide products indicates both the large and small polypeptides to be basic, the M_r = 38,000 polypeptide having a net excess of positively charged residues of +18 and the M_r = 25,000 polypeptide having a net excess of +11. This would be consistent with either interacting with DNA or RNA; but a biochemical role for neither has yet been established, despite the fact that the incC locus is known to express incompatibility, possibly through an effect on the level of trfA operon expression separate from that due to repression by the products of trfB and korB of the trfA promoter (23).

Transcriptional signals

The operon coding for trfB, incC and korB is transcribed from a promoter, the -35 region of which overlaps the HincII site (14). Inverted repeats I and II (Fig. 2) represent the putative binding sites for KorB (the korB gene product) and TrfB proteins respectively, both of which are transcriptional repressors of this operon (12,14,16). The DNA sequence presented here reveals that this promoter is preceded by a possible rho-independent transcriptional terminator (Fig. 2). The nature of the genes or transcriptional units which this terminator might be related to are at present unknown.

Evolution of the overlap between incC and trfB

It remains to be determined what purpose the overlap between incC and trfB serves. However, it seems likely that originally this region coded for just one gene and that it has subsequently gained the second functional ORF. In such a situation it has been proposed (for a review see ref. 32) that one should be able to determine which gene is the more ancient by applying an analysis which was originally designed to allow one to distinguish which of the three possible phases is most likely to represent a translated ORF (35). It has been proposed that the original genetic code was 'comma-free' and utilized only a subset of the 64 possible

triplets so that only one of the three phases would contain a series of such triplets that could be translated (36). The currently favoured proposal is that the original code consisted of RNY triplets (R ≡ purines; N ≡ any base; Y ≡ pyrimidines) (37,35). Calculation of the minimum number of point mutations required to return each phase to a series of RNY triplets has been successful in predicting all known genes in, for example the Øx174 genome (35) except in regions of gene overlap where inevitably only one of the two genes is identified. When such analysis is applied to the region of overlap between trfB and incC it is interesting to discover that the incC ORF not only has the greater number of RNY codons (37 compared to 19 in the trfB ORF) but also requires the smaller number of point mutations to return the other codons to RNY (84 compared to 108 for the trfB ORF). This would suggest that incC is the older gene in this region; a rather surprising conclusion in view of the central role of the trfB gene product in regulation of operons coding for replication and maintenance functions of RK2 (12, 15, 17, 18, 22).

One must therefore consider the possibility that originally the trfB operon contained only incC and korB and that regulation of transcriptional initiation in the kil and trfA operons was mediated only by the korB product. This proposal seems plausible since all the operons which have so far been investigated in this coordinately regulated system, trfA, trfB, kilA and putative kilB (12, 15,17,18) are repressed by the korB product, while only trfA, trfB and kilA operons are repressed by the trfB product. Therefore korB is a more general regulatory gene than trfB. In addition, the similarity of the putative trfB and korB protein binding site suggests that the trfB operator sites could have evolved from korB operator sites (or vice versa). Further analysis of this coordinately regulated series of operons on RK2 or on one distantly related IncP plasmid currently underway in this laboratory may provide further information relevant to this issue.

ACKNOWLEDGEMENTS

We thank R. Spooner for advice on use of the DNA sequence analysis programme; G. Ware of the Microbiology Department, Bristol University for permission to use his DNA sequence analysis computer programme; E. Badger for typing the manuscript and B.S. Price for help in preparing the Figures. C.A. Smith was supported by and this research was partly financed by MRC Project Grant G80/0374/9CB and G82/2421/3CB. Current support for this work is from MRC Project Grant G83/0983/8CB.

*To whom correspondence should be addressed

REFERENCES

1. Thomas, C.M. (1981) Plasmid 5, 10-19.
2. Thomas, C.M., Meyer, R. and Helinski, D.R. (1980) J. Bacteriol. 141, 213-222.
3. Meyer, R. and Helinski, D.R. (1977) Biochim. Biophys. Acta 487, 109-113.
4. Figurski, D. and Helinski, D.R. (1979) Proc. Natn. Acad. Sci. USA 76, 1648-1652.
5. Thomas, C.M. (1981) Plasmid 5, 277-291.
6. Cross, M.A., Warne, S.R. and Thomas, C.M. (1986) Plasmid, (in press).
7. Schmidhauser, T.J., Filutowicz, M. and Helinski, D.R. (1983) Plasmid 9, 235-330.
8. Schmidhauser, T.J. and Helinski, D.R. (1985) J. Bacteriol. 164, 446-445.
9. Smith, C.A. and Thomas, C.M. (1983) Mol. Gen. Genet. 190, 245-254.
10. Shingler, V. and Thomas, C.M. (1984) J. Mol. Biol. 175, 229-250.
11. Smith, C.A. and Thomas, C.M. (1984) J. Mol. Biol. 175, 251-262.
12. Shingler, V. and Thomas, C.M. (1984) Mol. Gen. Genet. 195, 523-529.
13. Pohlman, R.F. and Figurski, D.H. (1983) J. Bacteriol 156, 584-591.
14. Figurski, D.H., Pohlman, R.F., Bechhofer, D.H., Prince, A.S. and Kelton, C.A. (1982) Proc. Natn. Acad. Sci. USA 79, 1935-1939.
15. Smith, C.A., Shingler, V. and Thomas, C.M. (1984) Nucleic Acids Res. 12, 3619-3630.
16. Smith, C.A. and Thomas, C.M. (1984) J. Gen. Microbiol. 130, 1651-1663.
17. Theophilus, B.D.M., Cross, M.A., Smith, C.A. and Thomas, C.M. (1985) Nucleic Acids Res. 13, 8129-8142.
18. Young, C., Prince, A.S. and Figurski, D.H. (1985) Proc. Natn. Acad. Sci. USA 82, 7374-7378.
19. Young, C., Bechhofer, D.H. and Figurski, D.H. (1984) J. Bacteriol. 157, 247-252.
20. Thomas, C.M., Smith, C.A., Shingler, V., Cross, M.A., Hussain, A.A.K. and Pinkney, M. (1985) in 'Plasmids in Bacteria', Helinski, D.R., Cohen, S.N., Clewell, D.B., Jackson, D.A. and Hollaender, A., Eds., pp. 261-276, Plenum, New York/London.
21. Meyer, R. and Hinds, M. (1982) J. Bacteriol. 152, 1078-1090.
22. Thomas, C.M. and Hussain, A.A.K. (1984) EMBO J. 3, 1513-1519.
23. Thomas, C.M. (1986) Plasmid, (in press).
24. Bechhofer, D.H. and Figurski, D.H. (1983) Nucleic Acids Res. 11, 7453-7469.
25. Sancar, A., Hack, A.M. and Rupp, W.D. (1980) J. Bacteriol. 137, 692-693.
26. Kahn, M., Kolter, R., Thomas, C., Figurski, D., Meyer, R., Remaut, E. and Helinski, D.R. (1979) in 'Methods in Enzymology Vol. 68', Wu, R. Ed., pp. 268-280. Academic Press, New York.
27. Birnboim, H.C. and Doly, J. (1979) Nucleic Acids Res. 7, 1513-1523.
28. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) 'Molecular Cloning A Laboratory Manual', Cold Spring Harbor.
29. Lathe, R., Skory, S. and Kieny, M.P. (1984) Focus (BRL) 6, #4.
30. Ubben, D. and Schmitt, R. (1986) Gene, (in press).
31. Maxam, A.M. and Gilbert, W. (1980) in Methods in Enzymology Vol. 65, Grossman, L. and Moldave, K. Eds., pp. 499-560. Academic Press, New York.

32. Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F.P. and Olsson, O. (1983) Ann. Rev. Genet. 17, 499-525.
33. Tacon, W., Carey, N. and Emtage, S. (1980) Mol. Gen. Genet. 177, 427-438.
34. Pabo, C.O. and Sauer, R.T. (1984) Ann. Rev. Biochem. 53, 293-321.
35. Shepherd, J.C.W. (1981) Proc. Natn. Acad. Sci. USA 78, 1596-1600.
36. Crick, F.H.C., Brenner, S., Klug, A. and Pieccznik, G. (1976) Origins of Life 7, 389-397.
37. Eigen, M. (1978) Naturwissenschaften 65, 341-369.
38. Norrander, J., Kempe, T. and Messing, J. (1983) Gene 26, 101-106.