Complete sequence of pSC101

Alberto Bernardi and Françoise Bernardi

Laboratoire d'Enzymologie du C.N.R.S., 91190 Gif-sur-Yvette, France

Received 23 November 1984; Accepted 29 November 1984

ABSTRACT

We have completed the pSC101 sequence. The coding capacities of the newly sequenced regions show the presence of two large open reading frames close to the *oriT* region. Their size and localization suggest that these polypeptide chains could be involved in the transfer process of pSC101.

INTRODUCTION

pSC101 is a low copy number plasmid isolated from *Salmonella* (1), which confers tetracycline resistance to the recipient cell and has been the first cloning vector used for in vitro recombinant DNA experiments (2). However, because of its unique cloning site and the lack of phenotypical criterion to detect the presence of an inserted fragment, it was rapidly superseded by others, more convenient vectors with multiple cloning sites and allowing an easy screening. Furthermore, these cloning vectors having higher copy number give a better yield of cloned DNA fragments. Nevertheless, low copy number plasmids are still very useful vectors, for instance when the overexpression of a given gene is deleterious to the host.

Much work is still devoted to pSC101 in several laboratories essentially because of its stringent mode of replication and its absolute requirement of *dnaA* gene product for replication (3). This requirement shared by the *oriC* region of *E. coli*, makes of this plasmid the best available model to study *E. coli* replication.

The presence of two insertion elements, IS101 (4) and IS102 (5-6) has been revealed by the biological activities normally displayed by these elements and this makes a further reason of interest in pSC101. In the studies we are performing on atypical deletions generated by IS102 elements modified in vitro (7) we have observed that frequently the endpoints of the deletions are not located at the end of the element but in pSC101 regions that we cannot localize precisely because of the lack of sequence data.

Actually, in spite of its relatively small size, the complete sequence of pSC101 has not yet been worked out and only partial sequence data scattered in several papers are available. The knowledge of the complete sequence of this plasmid would allow a precise localization of mutants and a better vector could be constructed. For these reasons we have thought to be of interest to many laboratories studying different aspects of pSC101 to have the complete sequence of this plasmid.

In this communication the complete sequence of pSC101 is presented: part of it is taken from the literature, proper references are provided, and part is our recent results.

MATERIALS AND METHODS

DNA sequencing method

DNA sequence was essentially determined according to Sanger (8) except for the use of deoxyadenosine 5' ³⁵S α (thio)triphosphate (9) and for the choice of the cloning system which in our case was the vector pEMBL8(+) (10). The sequence of the complementary strand, or the opposite end of a given fragment in case of fragments too long to be sequenced from one end to the other, was occasionally determined by cross-hybridizing several independent clones obtained in the same experiment; more frequently we have preferred the following procedure: plasmid DNA obtained from a small culture (11) was degraded with EcoRI and HindIII which flank the polylinker region; the fragment so obtained was purified on an acrylamide gel and subcloned in the pEMBL9(+) vector (10) degraded with the same restriction enzymes. Because of the asymmetrical sequence of the ends of the fragment and of the vector and of the opposite orientation of the corresponding polylinker region in the vector only clones having the opposite orientation as compared to the primer region can be obtained. All the sequence data we present have been obtained by cloning in the HindII site after having linearized the vector with HindII or AccI when HpaII fragments were to be cloned. The single stranded DNA was prepared by superinfection with IRI phage (10) and the sequence was determined using the pentadecamer primer. Some clones which reproducibly gave an ambiguous gel reading were sequenced directly on double stranded DNA (12). In this case plasmid DNA containing the fragment of interest was degraded by PvuII which in pEMBL has two restriction sites flanking the polylinker region (10); this fragment was fractionated on acrylamide gel and directly sequenced. A few sequence data were obtained according to the Maxam Gilbert method (13).

The pentadecamer primer and the dideoxynucleotides mixtures were from Biolabs or they were prepared according the same composition, with minor adjustments, with Boehringer products.

Enzymes

Klenow fragment was obtained from Biolabs or from Boehringer or else was prepared from the strain CJ155 (14) kindly provided by Dr. Joyce. DdeI, PvuII, AccI, were from Biolabs; the other restriction enzymes were prepared according to Ref. 15. S1 nuclease was from Boehringer.

RESULTS and DISCUSSION

General strategy

We have chosen to use the recently described pEMBL vectors because the possibility of DNA arrangements is less likely to occur with this vector due to the fact that the single strand replication mode is only used for a very limited number of duplications. To complete the sequence of pSC101 fragments 2 and 3 (fragments are numbered in order of decreasing molecular weight after an HindII degradation) needed to be sequenced entirely while large sequence of fragments 1 and 6 are already available. Therefore we have cloned independently these four fragments in the unique HindII site of the vector pEMBL8(+). Next, each clone was examined after degradation with AluI, BspRI (16) an isoschizomer of HaeIII, HpaII and TaqI in parallel with the vector carrying no insert and the bands of pSC101 origin were identified; they were eluted from the gel, subcloned in the same vector linearized with the appropriate restriction enzyme and sequenced. The sequence of the complementary strand was determined by subcloning the fragment in the vector pEMBL9 as described in Materials and Methods. As all the data obtained in this way did not allow to reconstitute the complete sequence and to confirm the sequence from independent clones, additional data were also obtained by studying some inserts with HinfI and DdeI. In this case prior to cloning, the sticky ends of the fragments were first trimmed with S1 nuclease. The extent of the sequence in each experiment as well as the restriction sites we have used (only as far as the sequence of the larger region we have sequenced is concerned) are presented in Fig. 1. About 90% of this fragment has been sequenced on both strands and because of the particular strategy used which generates redundant data, many regions were sequenced more than once on the same strand. These informations, however are not always reported in Fig. 1. The complete sequence formed by 9263 bp was obtained and is presented in Fig. 2. The sequence starts at the unique

Fig. 1: Strategy used to determine the sequence of part of *HindII* fragment 6 (located on the left side) and the adjacent fragments 2 and 3. Only the restriction sites used for the sequence work are indicated. The same coordinates as in Fig. 2 are indicated in the lowest line.

EcoRI site and the same numbering as presented by Sutcliffe (17) is followed in the first part.

Organisation and sources of the sequences

The sequence presented in Fig. 2 has been obtained from position 1 to position 1643 (comprising the tetracycline gene) as determined by Sutcliffe (17) on pBR322. We have inserted at position 529 an additional C residue as shown by Peden (18) which gives an excellent agreement between the molecular weight of the tetracycline gene product as determined from the DNA sequence and that experimentally observed. At position 1643 the pSC101 sequence diverges from that of pBR322. We have been surprised not to find an *EcoRI** restriction site (19) (NAATTN) at this junction. Such a site, according to the construction (20) should actually be present both in pSC101 and in pBR322 while it is absent from both plasmids. According to our sequence it is likely that in this particular case *EcoRI** has recognized a sequence similar to *EcoRI* site but having one mismatch namely GAATGC. From position 1643 to position 4510 is the sequence we have determined and whose map is reported in Fig. 1. IS101 is located in this region (from position 2024 to position 2232); we have redetermined its complete sequence and found no difference with the previously published one (4). The unique *nic* site present on pSC101, which is part of the origin of the transfer of this plasmid has been mapped in this region also (see below). The sequence of the *par* region and of the replication origin has been determined in several laboratories (21-24) and there is a very good agreement between the proposed sequences. Here we report the sequence (numbered in Fig. 2 from position 4511 to position 6576) presented by Armstrong et al. (24) because it

TTCTCATGTT	TGACAGCTTA	TCATCGATAA	GCTTTAATGC	GGTAGTTTAT	CACAGTTAAA	60
TTGCTAACGC	AGTCAGGCGC	CGTGTATGAA	ATCTAACAA	GCCTCATCG	TCATCCTCGG	120
CACCGTACCC	CTGGATGCTG	TAGGCATAGG	CTTGTTATG	CCGGTACTGC	CGGGCCTCTT	180
GCGGGATATC	TGCCATTCCG	ACAGCATCGC	CAGTCACTAT	GGCGTGTGC	TAGCGCTATA	240
TGCGTTGATG	CAATTTCTAT	GCGCACCCGT	TCTCGGAGCA	CTGTCCGACC	GCTTTGGCCG	300
CCGCCACGTC	CTGCTCGCTT	CGCTACTTGG	AGCCACTATC	GACTACGCGA	TCATGGCGAC	360
CACACCCGTC	CTGTGGATCC	TCTACGCCGG	ACGCATCGTG	GCCGGCATCA	CCGGCGCCAC	420
AGGTGCGGTT	GCTGGCGCCT	ATATCGCCGA	CATCACCAGT	GGGGAAGATC	GGGCTCGCCA	480
CTTCGGGCTC	ATGAGCGCTT	GTTCGGGCGT	GGGTATGGTG	GCAGGCCCCC	TGGCCGGGGG	540
ACTGTTGGGC	GCCATCTCCT	TGCATGCACC	ATTCTTGGG	CGGGCGGTGC	TCAACGGCCT	600
CAACCTACTA	CTGGGCTGCT	TCCTAATGCA	GGAGTCGCAT	AAGGGAGAGC	GTGACCCGAT	660
GCCCTTGAGA	GCCTTCAACC	CAGTCAGCTC	CTTCCGGTGG	GCGCGGGGCA	TGACTATCGT	720
CGCCGCACTT	ATGACTGTCT	TCTTTATCAT	GCAACTCGTA	GGACAGGTGC	CGGCAGCGCT	780
CTGGGTCATT	TTCCGGCAGG	ACCGCTTTCG	CTGGAGCGCG	ACGATGATCG	GCTGTGCGCT	840
TGCGGTATTC	GGAACTTTGC	ACGCCCTCGC	TCAAGCCTTC	GTCACTGGTC	CCGCCACCAA	900
ACGTTTGGCG	GAGAGCAGGG	CCATTATCGC	CGGCATGGCG	GCCGACGCGC	TGGGCTACGT	960
CTTGTGCGCG	TTCCGCGACG	GAGGCTGGAT	GGCTTTCGCC	ATTATGATTC	TCTCGCTTC	1020
CGGCGGCATC	GGGATGCCCG	CGTTGCAGGC	CATGCTGTCC	AGGCAGGTAG	ATGACGACCA	1080
TCAGGGACAG	CTTCAAGGAT	CGCTCGCGGC	TCTTACCAGC	CTAACTTCGA	TCACTGGACC	1140
GCTGATCGTC	ACGGCGATTT	ATGCCGCTCT	GGCGAGCACA	TGGAACGGGT	TGGCATGGAT	1200
TGTAGGCGCC	GCCCTATACC	TTGTCTGCGT	CCCCGCGTTG	CTCTCGGGTG	CATGGAGCCG	1260
GGCCACCTCG	ACCTGAATGG	AAGCCGGCGG	CACCTCGCTA	ACGGATTAC	CACTCCAAGA	1320
ATTGGAGCCA	ATCAATTCTT	CGCGAGAACT	GTGAATGCGC	AAACCAACCC	TTGGCAGAGA	1380
ATATCCATCG	CGTCCGCCAT	CTCCAGCAGC	CGCACGCGGC	GCATCTCGGG	CAGCGTTGGG	1440
TCCTGGCCAC	GGGTGCGCAT	GATCGTGCTC	CTGTCTTGA	GGACCCGGCT	AGGCTGGCGG	1500
GGTTGCGCTTA	CTGGTTAGCA	GAATGAATCA	CCGATACGCG	AGCGAAGCTG	ACAGCATGCG	1560
TGCTGCAAAA	CGTCTGGAC	CTGAGCAACA	ACATGAATGG	TCTTCGGTTT	CGCTGTTTTC	1620
TAAAGTCTGG	AAACGCGGAA	GTCCCTACG	TGCTGCTGAA	GTTCGCCGCA	ACAGAGAGTG	1680
GAACCAACCC	GTGATACCAC	GATACTATGA	CTGAGAGTCA	ACGCCATGAG	CGGCCTCATT	1740
TCTTATCTG	AGTTACAACA	GTCCGCACCC	CTGCCGGTAG	CTACTTGACT	ATCCGGCTGC	1800
ACTAGCCCTG	CGTCAGATGG	CTGTGATCCA	AGGCAAACTG	CCAAATATATC	TGCTGGCACC	1860
GGAGTCAAGC	GCCCTGCACC	ATTATGTTCC	GGATCTGCAT	CGCAGGATGC	TGCTGGCTAC	1920
CCTGTGGAAC	ACCTACATCT	GTATTAAACGA	AGCGCTGGCA	TTGACCTTGA	GTGATTTTTC	1980
TCTGGTGCCG	CCCTATCCCT	TTGTGCAGCT	TGCCACGCTC	AAAGGGGTTT	GAGGTTCAAC	2040
CGTACGAAAA	CGTACGGTAA	GAGGAAATTT	ATCGTCTGAA	AAATCGATTA	GTAGACAAGA	2100
AAGTCCGTTA	AGTGCCAAAT	TTCGATTAA	AAGACACCGT	TTTGATGGCG	TTTTCCAATG	2160
TACATTATGT	TCGATATAT	CAGACAGTTA	CTTCACTAAC	GTTACGTTTTC	GTCTTATTTG	2220
CCTTCAGACC	CCATCCCTTT	AATGTCCTTT	ATTTGCTGGG	GTATTCAGAT	CCCCCGGACA	2280
CGTTTAATTA	ATGCTTTCCTC	CGCCGGAGAT	CGACGCACAG	GCTTCTGTGT	CTATGATGTT	2340
ATTTCTTAAT	AATCATCCAG	GTATTCTCTT	TATCACCATA	CGTAGTGCGA	GTGTCCACCT	2400
TAAACGAGGG	CTTTCGCTCA	CAGCGCGATA	TGTCAGCCAG	CGGGGCTTTT	TTTTGCCAGA	2460
CCGCTTCCAT	CCTCTGCATT	TCAGCAATCT	GGCTATACCC	GTCATTCTATA	AACCACGTAA	2520
ATGCCGTCAC	GCAGGAGATC	AGGACGAAGA	ATATCGTCAG	TACAGATATA	ATCGCGGATT	2580
TCCACGTATA	CGCTGACATC	TCACGACGCA	TTTCATGGAT	CATCGCTTTC	CGCGTATCGG	2640
CAGCCTGATT	CAGCGCTTCT	GTGCGCGGTT	TCTGCTGTGC	TAATCCGGCT	TGTTTTCAGTT	2700
CTTTCTCAAC	CTGAGTGAGC	GCGGGACTCA	CCGATTTCCCT	GACGGGTGTA	GTCAATATTAC	2760
CGGACGCGCT	GCTCAGCTCA	CGAATGACCC	TGCTCAGCGT	TTCACTTTGC	TGCTGTAAAT	2820
GTGATGAGGC	GGCTTGAATC	TGTTCTGTCA	GAGAAGTAAC	ACGCTTTTCC	AGCGCTGTAT	2880
GATGCCCGAT	AAGGGCGGCA	ATTTGTTTAA	TTTCGTGCGT	CATACAAAAT	CCTGCTATC	2940
GTGAGAATGA	CCAGCCTTTA	TCCGGCTTCT	GTCTGATCTG	TTCCGGCGAGT	CGCTGTCTGT	3000
CTTTCTCCTG	CTGACGCTGT	TTTTCCGCCA	GACGTTCCGCG	CTCTCTCTGC	CTTTCCATCT	3060
CCTGATGTAT	CCCCTGGAAC	TCCGCCATCG	CATCGTTAAC	AAGGGACTGA	AGATCGATTT	3120
CTTCTCTGAT	ATCCTTTCATG	GCATCACTGA	CCAGTGCGTT	CAGCTTGTC	GGCTCTTTT	3180
CAAAATCAAA	CGTCTGCCC	GAATGGGAT	CCTGCTCAGG	CTCTGCTTTC	TGCGCTGTT	3240
TTAGCGTCAG	AGTATCCCTC	TCGCTGAGGG	CTTCCCGTAA	CGAGGTAGTC	ACGTCAATTA	3300
CGCTGTCAG	TTCACTACGG	GACTGCTGCA	CCTGCTTTTC	AGCCTCCCTG	CGCTCAGGAA	3360
TGGCCTGTAG	ATGCTCAGTA	TGCAATCGCT	GAACCTGACC	CGCGCCAGAG	TGCGCTCAG	3420
GCTCACGGTC	CATGCCCTGC	GCCTTCAGGG	AACGGGAATC	AACCCGGTCA	CGGTGCTGAT	3480
ACCGTTCAAG	GTGCTTATTC	TGGAGGTCAG	CCAGCGCTCT	CCCTCTGGGC	AACACGGGAT	3540
TCTTTGCGTT	CGGTGCGGTG	TTCCCCGAAA	CGTGCTTTT	TTGCGCCACC	CGGTCCGGCT	3600
CTTTGCTGTT	AGCCCGTTTA	AAATAGTGCT	CAGGGTCACG	GTGAATACCG	TCATTAAATGC	3660

GTTCAGAGAA	CATGATATGG	GCCTGGGGCT	GCTCGCCACC	GGCTATCGCT	GCTTTCGGAT	3720
TATGGATAGC	GAACCTGATG	GCATGGCGGT	CGCCAAATTC	CTGTTGGACA	AAATCGCGGA	3780
CAAGCTCAAG	ACGTTGTTCC	GGTTTTAACT	CACGCGGCGA	GGCAATCTCG	ATTTACACGGT	3840
AGGTACAGCC	GTGGACAGT	TCAGACGTGT	CAGCGGCTTT	CCAGAACTCG	GACGGTTTAT	3900
GCCTGCCCCA	CGCCGGCATA	TTGCCGGACT	CCTTGTGCTC	ACAGGTCGGAG	TCTTTTTCAC	3960
GGGCATACTT	TCCTCTACGC	GCAATATAAT	CGGCATGAGG	AGAGGCACTG	CCTTTTCCGC	4020
CGGTTTTTAC	GCTGAGATGA	TAGGATGCCA	TCGTGTTTTA	TCCCCTGTA	GGGCGCACGT	4080
TTCTGAACGA	AGTGAAGAAA	GTCTAAGTGC	GCCCTGATAA	ATAAAGAGT	TATCAGGGAT	4140
TGTAGTGGGA	TTTGACCTCC	TCTGCCATCA	TGAGCGTAAT	CATTCCGTTA	GCATTCAGGA	4200
GGTAACAGCG	ATGAATAAAA	GCAGAAAAAC	AGGAACAATG	GGCAGCAGAA	AGAGTGCAGT	4260
ATATTCCGCG	CTTAAGTCCG	CCGAATGAGC	AACAGAAACT	TATGCTGATA	CTGACGGATA	4320
AAGCAGATAA	AACAGCACAG	GATATCAAAA	CGCTGTCCCT	GCTGATGAAG	GCTGAACAGG	4380
CAGCAGAGAA	AGCGCAGGAA	GCCAGAGCGA	AAGTCATGAA	CCTGATACAG	GCAGAAAAAGC	4440
GAGCCGAGAG	CAGAGCCGCC	CGTAAAGCCC	GTGACCATGC	TCTGTACCA	TCTGCCGGAT	4500
TGCTTATCCT	GGCGGGTCTG	GTTGACAGTA	AGACGGGTAA	GCCTGTTGAT	GATACCCGCTG	4560
CCTTACTGGG	TGCATTAGCC	AGTCTGAATG	ACCTGTCACG	GGATAATCCG	AAGTGGTCAG	4620
ACTGGAATAA	CAGAGGGCAG	GAACCTGCTG	ACAGCAAAAA	GTCAGATAGC	ACCACATAGC	4680
AGACCCGCCA	TAAACGCCCC	TGAGAAGCCC	GTGACGGGCT	TTCTTTGAT	TATGGGTAGT	4740
TTCTTTGCAT	GAATCCATAA	AAGGCGCCTG	TAGTGCCATT	TACCCCATTT	CACTGCCAGA	4800
GCCGTGAGCG	CAGCGAACTG	AATGTCACGA	AAAAGACAGC	GACTCAGGTT	CCTGATGGTC	4860
GGAGACAAAA	GGATATTCCA	GCGATTTGCC	CGAGCTTGCG	AGGGTGCTAC	TTAAGCCTTT	4920
AGGGTTTTAA	GGTCTGTTTT	GTAGAGGAGC	AAACAGCGTT	TGCGACATCC	TTTTGTAAATA	4980
CTGCGGAAGT	GACTAAAGTA	GTGAGTTATA	CACAGGGCTG	GGATCTATTC	TTTTTATCTT	5040
TTTTTATCTT	TTCTTTATTC	TATAAATTAT	AACCACTTGA	ATATAAACAA	AAAAAACACA	5100
CAAAAGGCTA	CGGGAATTTA	AGCAGAGTCT	AGCAGAAATT	ACAAGTTTTT	CAGCAAGGTT	5160
CTAGCAGAAAT	TTACAGATAC	CCACAACCTA	AAGGAAAAGG	ACTAGTAATT	ATCATTGACT	5220
AGCCCATCTC	AATTGGTATA	GTGATTAAAA	TCACCTAGAC	CAATTGAGAT	GTATGTCCTGA	5280
ATTAGTTGTT	TTCAAGACAA	GATTAGTCGC	GATTAGTCGC	TATGACTTAA	CGGAGCATGA	5340
AACCAAGCTA	ATTTTATGCT	GTGTGGCACT	ACTCAACCCC	ACGATTGAAA	ACCCTACAGG	5400
GAAAGAACGG	ACGGTATCGT	TCACCTATAA	CCAATACGCT	CAGATGATGA	ACATCAGTAG	5460
GGAAATGCT	TATGGTGTAT	TAGCTAAAGC	AACCAAGAGG	CTGATGACGA	GAGCTGTGGA	5520
AATCAGGAAT	CCTTTTGGTTA	AAGGCTTTGA	GATTTTCCAG	TGGACAAACT	ATGCCAAGTT	5580
CTCAGCGGAA	AAATTAGAAAT	TAGTTTTTAG	TGAAGAGATA	TTGCCTTATC	TTTTCCAGTT	5640
AAAAAAATCT	ATAAAGTATA	ATCTGGAACA	TGTTAAGTCT	TTTGAACACA	AATACTCTAT	5700
GAGGATTTAT	GAGGATTTAT	TAAAGAAACT	AACACAAAAG	AAAAGCTACA	AGGCAATAT	5760
AGAGATTAGC	CTTGATGAAT	TTAAGTTCAT	GTAAATGCTT	GAAAATAACT	ACCATGAGTT	5820
TAAAGGGCTT	ATGCCAATGGG	TTTTGAACCT	AATAAGTAAA	GATTTTAAACA	CTTACAGCAA	5880
TATGAATTTG	GTGGTTGATA	AGCGAGGCCG	CCCGACTGAT	ACGTTGATTT	TCCAAGTTGA	5940
ACTAGATAGA	CAAATGGATC	TCGTAACCGA	ACTTGAGAAC	AACCAAGATA	AAATGAATGG	6000
TGACAAAATA	CCAACAACCA	TTACATCAGA	TTCTACCTA	CGTAACGGAC	TAAGAAAAAC	6060
ACTACACGAT	GCTTTAACTG	CAAAAATTCA	GCTCACCAGT	TTTGAGGCAA	AAATTTTGAG	6120
TGACATGCAG	AGTAAGCATG	ATCTCAATGG	TTGCTTCTCA	TGGCTCACGC	AGGAACACAG	6180
AACCACTACT	GAGAACATAC	TGGCTAAATA	CGGAAGGATC	TGAGGTTCTT	ATGGCTCTTG	6240
TATCTATCAG	TGAAGCATCA	AGACTAACAA	ACAAAAGTAG	AACAAGTGT	CACCGTTAGA	6300
TATCAAGGGG	AAAAGTGTCC	ATATGCACAG	ATGAAAACGG	TGTAAAAAAG	ATAGATTACAT	6360
CAGAGCTTTT	ACGAGTTTTT	GGTGCATTTA	AAGCTGTTCA	CCATGAACAG	ATCGACAATG	6420
TAAACAGATGA	ACAGCATGTA	ACACCTAATA	GAAACAGGTA	AACCAAGTAA	ACAAGCAAC	6480
TAGAACATGA	AATTGAACAC	CTGAGACAAC	TTGTTACAGC	TCAACAGTCA	CACATAGACA	6540
GCCTGAACAA	GGCGATGCTG	CTTATCGAAT	CAAGCTGCC	GACAACACGG	GAGCCAGTGA	6600
CGCTCCCGCT	GGGGAAAAAA	TCATGGCAAT	TCTGGAAGAA	ATAGCGCTTT	CAGCCGGCAA	6660
ACCTGAAGCC	GGATCTGCGA	TTCTGATAAC	AAACTAGCAA	CACCAAGACA	GCCCCTTTCG	6720
GGGACGCAAA	ACCCGTACTT	TGTGACGTTT	CGGCGGTTTT	TTGTGGCGAG	TGGTGTTCGG	6780
GCGGTGCGCG	CAAGATCCAT	TATGTTAAAC	GGGCGAGTTT	ACATCTCAAA	ACCGCCCGCT	6840
TAAACCATCT	AGAAATCCTC	AGCGCGATTT	TAAAGACCAA	CCCCCGCCCC	TAAACCCCAA	6900
ATCCATCATG	AGAAGTGGCT	TGTTGAATAA	ATCGAAGTTT	GCTTGGCTTG	AGGATCAGAA	6960
TCACGCATCC	TCCCGACAAC	ACAGACCATT	CCGTGGCAAA	GCAAAAGTTC	AGAATCACC	7020
ACTGGTCCAC	CTACACACAA	GCTCTCATCA	ACCGTGGCTC	CCTCACTTTT	TGGCTGGATG	7080
ATGAGGCGAT	TAGGGCTTGG	TATGAGTCGG	CAACACCTTC	ATCACGAGGA	AGGCCCCAGC	7140
GCTATTCTGA	TCTCGCCATC	ACCACCGTTC	TGGTGAATTA	ACGCGTATTC	CGGCTGACCC	7200
TGCGGGGCTG	GCAGGGTTTT	ATTGATTCCA	TTTTTGCCCT	GATGAACGTT	CCGTTGCGCT	7260
GCCCGGATTA	CACCAAGTGC	AGTAAGCGGG	CAAGTCGGT	TAATGTCAGT	TTCAAAACGT	7320

```

CCACCCGGGG  TGAATCGCA  CACCTGGTGA  TTGATTCCAC  CGGGCTGAAG  GTCTTTGGTG  7380
AAGGCGAATG  GAAAGTCAGA  AAGCACGGCA  AAGAGCGCCG  TCSTATCTGG  CGAAAGTTGC  7440
ATCTTGCTGT  TGACAGCAAC  ACACATGAAG  TTGTCTGTGC  AGACCTGTGC  CTGAATAACG  7500
TCACGGGACT  AGAAGCCTTC  CCGGGCCTTA  TCCGGCAGAC  TCACAGAAAA  ATCAGGGCAG  7560
CCGCGGCGAG  CGGGGCTTAC  GATACCCGGC  TCTGTACGSA  TGAAGTGCAG  CGCAAAAAAA  7620
TCAGCGCGCT  TATTCCTCCC  CGAAAGGGTG  CGGGTTACTG  GCCCCTGTAA  TATGACAGCC  7680
GTAAACGTGC  AGTGGCTAAT  CAGCGAATGA  CCGGGAGTAA  TGCCTGGTGG  AAATGGACAA  7740
CAGATTACAA  CCGTGCCTCG  ATAGCGGAAA  CGGCGATGTA  CCGGGTAAAA  CAGCTGTTCT  7800
GGGGTTCAAT  GACGCTGCGT  GACTACGATG  GTCAGGTTGC  GGAGGCTATG  GCCCTGGTAC  7860
GAGCGCTGAA  CAAATGACG  AAGCAGGTA  TGCTGAAAG  CGTGCCTATT  GCCTGAAAC  7920
ACACCCGCT  ACGGGGGAGA  CTTACCCGAA  ATCTGATTTA  TTCAACAAAG  CCGGGTGTGG  7980
TGAACTACAA  AGCAGACCCG  TTGAGGTTAT  CAGTTGATG  CACAATCAGC  AGCGCATAAA  8040
ATATGCACAA  GAACAGGAGC  ACCCTTCGCA  TTAAGCTGTG  GTGGTAACAA  GTAGTGCCGG  8100
GCTACCATCA  GCGAGCATGA  TGCCTCCCA  CAGCATTCGC  CTTGGCAGTA  TGGAGTTCC  8160
TCGCTCCAGT  TCGGGCCGGT  ATCCACCTCG  AGTGACCCCA  GCCGCCCTC  ATGCCAACCG  8220
AGCCCATTAA  GCGCAGAGTC  GGGCGCAACT  ATCTCGGGGA  CGCATAAACA  CCGCAGGAT  8280
TACCAAGGAG  TTCGGCTAGG  TCAATCCGCA  ATTCCCTTTC  CTGGCCGGA  GATGCCGCGA  8340
GCAACAATG  TGTCCGGTGC  CTACAGATG  GACTTTGTGG  GCGGAGGATT  GCACAGAGCG  8400
CTTCAGCGCA  GCGTCTGATG  ATGAATCGGA  AAGCCCGGTA  AGCAATTGAA  GAACCGCTGA  8460
CCACCTCCCA  TCCGACAATG  CAAGGGTGCC  CCGCTCCAG  CGCGACAGCG  TTGCTGTATC  8520
CACACCCATT  AATTCAGCTA  GGTGACTTTG  CTTCATATGA  CGAAGCAACC  GCGCCCTCCG  8580
CACCTGCCCG  CCCCTGTGAT  TCGTCAACAT  TCTTCAGCAC  CTCATGTGCG  TTCGTAGTGA  8640
GCCTCATTTT  TTCAAGACCG  CCGATGATGA  GAGCCAGGCC  TCGCTCGAAT  GCCGCGTCCG  8700
GACCGCCTTC  GATGACGATT  TTCATGCGCG  TCTGTAGGCG  CGCCGGCATC  GTAGAGCTCG  8760
AGGTGGTCAA  CTGATCTTCG  CCCCCTCCT  CCGCTCTGCG  CTCGCTAGCT  TGCTGCTCAA  8820
GAACAGCGCC  GACGGTGAAG  TAGCTGATTG  CCATCAACGC  ATAGGTCGCG  TCACCTGCCG  8880
AAAAGCCAGC  ATCGCAAGG  AAGCBAAGCT  GCGCTCTGCG  TTTTCCATC  TGGCGCGCGG  8940
CTGGCCGCGT  CCGGCGATGA  ATACGCGCGC  CATCGCGATA  AGCGAGCAAC  GCCCGTCGAA  9000
AACTGCATGC  ATTGCCCTTC  AGGAACGAAC  GCCAGTCGTC  GTCATCCCTT  GCGCTCGAAT  9060
GCGTGTGATT  TATGCTAGC  ATGGCTTCGG  CAAGTGCCTC  GAGCAACGCA  CGCTTGTCTC  9120
TGAATGACCA  GTAGAGCGCT  GGGTGTGCA  CCCCAGGCGC  CTCAGCCAGT  CCGCGCGCTCG  9180
TTAGACCTTC  CATGCCACG  TCGTTAAGCA  GTTCGAGCGC  GGTTCGGATC  ACBGCCTCGC  9240
GTTGGAGCTT  GTTCATTCGC  GAA  9263

```

Fig. 2: Complete sequence of pSC101. The sequence starts in the center of the unique EcoRI site and the numbering indicated by Sutcliffe (17) is followed in the first part. The sequence is written in the 5' to 3' order.

is the largest published sequence of this region. The sequence from this point to IS102 has been determined by us in this work and finally the sequence of IS102 (between positions 6916 and 7972) and the sequence comprising the repressor of the tetracycline gene is our previous work (6-7).

Restriction sites

Positions for several restriction sites on pSC101 are indicated in Table 1. Two StuI sites are found on pSC101 at position 7096 and 8679, however because the presence of a methylC at position 7097 overlapping an EcoRII (BstNI) restriction site, two fragments are obtained only when the plasmid is grown in a dcm⁻ strain. Similarly there are three ClaI sites at position 25, 2085 and 3115; however because of the presence of a methylA residue at position 3113, belonging to a MboI site, three fragments are obtained only in a dam⁻ strain.

TABLE 1. Restriction enzyme recognition sites in pSC101

Enzyme	Restriction sites ¹	Position in pSC101
<u>Acc</u> I	GT(A,C)(G,T)AC	653;2093;8713;8753
<u>Alu</u> I	AGCT	17;32;688;1091;1781;2009;2777;3164; 3233;3371;3785;4895;5348;5484;5501; 6092;6366;6394;6520;6576;7042;7794; 8076;8538;8809;8844;8909;9248.
<u>Ava</u> I	C(C,T)CG(A,G)G	1426;4890;7325;7521;8188;8254;9153.
<u>Ava</u> II	GG(A,T)CC	800;888;1137;1440;1482;2034;7025;8701.
<u>Bal</u> I	TGGCCA	1447.
<u>Bam</u> HI	GGATCC	376.
<u>Bsp</u> RI	GGCC	175;298;402;526;534;598;832;921;942; 993;1050;1263;1447;1734;2221;2833; 3364;5908;7096;7134;7526;7662;7852; 8176;8243;8326;8434;8679;8945;9235.
<u>Bst</u> NI	CC(A,T)GG	132;1061;1444;2359;2541;3075;4510; 7098;7345;7855;8323;8676.
<u>Cla</u> I	ATCGAT	25;2085;3115.
<u>Eco</u> RI	GAATTC	9262.
<u>Hae</u> II	(A,G)GCGC(C,T)	237;418;439;499;553;780;1210;1873; 1956;2657;2876;4768;6648;7143;7419; 7867;8830;9140;9162.
<u>Hind</u> II ²	GT(C,T)(A,G)AC	654;1720;3098;4524;7452;8606;8769
<u>Hind</u> III	AAGCTT	30.
<u>Hin</u> fI	GANTC	633;853;1007;1305;1526;1716;2648;2988; 3208;3384;3457;3928;3949;4752;4842; 5528;6030;6568;6680;7013;7089;7105; 7225;7354;7507;7539;8237;8424.
<u>Hpa</u> I	GTTAAC	3098
<u>Hpa</u> II	CCGG	162;171;388;403;412;535;695;771;931; 1021;1259;1285;1486;1689;1775;1794; 1860;1890;2304;2666;2686;2761;2963; 3199;3465;3596;3700;3914;3925;4021; 4496;6655;6670;6751;7191;7264;7326; 7361;7522;7533;7587;7664;7712;7782; 7972;8098;8177;8327;8355;8436;8699; 8744;8953.
<u>Mbo</u> II	GAAGA	477;731;1002;1594;2559;3113;3122;4107; 5625;6648;8461;8604;8768.
<u>Mlu</u> I	ACGCGT	7182.
<u>Pvu</u> II	CAGCTG	7794.
<u>Rsa</u> I	GTAC	166;2044;2054;2162;2203;2562;3845; 4486;6737;7780;7859;8418.
		.../...

(Table 1, continued)

Enzyme	Restriction sites ¹	Position in pSC101
<u>Sac</u> II	CCGCGG	7565.
<u>Sal</u> II	GTCGAC	652.
<u>Sma</u> II	CCCGGG	7327;7523.
<u>Stu</u> I	AGGCCT	7096;8679.
<u>Taq</u> I	TCGA	25;340;653;1128;1269;2085;2123;2173; 2311;3115;3382;3829;6413;6566;6933; 7759;8016;8189;8686;8997;9056;9100; 9214.
<u>Xho</u> I	CTCGAG	8188

No restriction sites for: ApaI, BclI, BglII, KpnI, PstI, SacI.

¹) The indicated positions correspond to the cleavage site.

²) Note that on the contrary to all previously published HindII maps fragment 2 (and not fragment 3) is adjacent to the tetracycline gene.

Sequences and coding capacities around the origin of transfer (oriT) of pSC101

pSC101 is a nonconjugative plasmid but mobilizable (mob⁺) at variable frequency according to the R-factors used for the transfer (25-27). Mobilization requires a cis-acting locus, oriT, where a nick occurs during mobilization, allowing the formation of a relaxation complex; plasmid encoded protein(s) (28) are also necessary in addition to the mobilizing R-factor. pSC101 nic region has been localized (29) to a 575 bp HinfI-HindII fragment (positions 3949-4524), close to the origin of vegetative replication. In this short region, we have found significant sequence homologies. Two direct repeats of 25 bp having 80% of homology (19 out 25 bp), start at positions 4237 and 4375 respectively. A region of twofold symmetry of 46 bp, where 34 are matched (Fig. 3) is located between positions 4070 and 4115. The presence of direct and inverted repeats close to each other has been also found in the oriT region of the plasmid RK2 (30). In addition there is an homology of 80% over 15 bp between the sequence at position 4354 and the oriT region of the plasmid RK2.

It has been previously shown that the nic region of colE1 and pBR322 have a striking homology with a region proximal to oriT of the F factor where 9 out 10 nucleotides are identical and it has been proposed that this sequence plays some role in the F promoted transfer of colE1 and pBR322

Analysis of the coding capacities of pSC101 shows the presence of two open reading frames in close proximity of oriT region between positions 4051 and 2938 (371 aminoacids) and between position 2923 and 2341 (194 aminoacids) respectively. In both cases the initiator codon is preceded by a possible Shine-Dalgarno sequence. The molecular weight of the largest polypeptide would be of 41 Kilodaltons, very close therefore to that of the tetracycline gene product and could have remained unnoticed in minicells extract. These polypeptides could be involved in the formation of the relaxation complex of pSC101. A similar organization has been described for ColE1 (28 and 31) where the oriT region is in proximity of a large mob region coding for the mobilizing protein(s) which are involved in the formation of the relaxation complex. The coding region of pSC101 we describe here has exactly the same size and position relative to the nic region as in colE1: we propose that it has also the same functions.

We wish to thank Drs. L. Dente and A. Parmeggiani for the strains used during the sequence work, Dr. F. Sor for having kindly provided the computers programs used for sequence analysis, Drs. C. Joyce and F. van

Horen for having provided the strains CJ155 and GM118 dcm⁻ and GM33 dam⁻ respectively, Dr. K. Armstrong for a preprint of the article quoted in Ref. 24 and Mrs J. Mauger for typing the manuscript. This work has been supported by grants from the "Fondation pour la Recherche Médicale Française" and from the "Délégation Générale à la Recherche Scientifique et Technique" to A.B.

REFERENCES

1. Cohen, S.N. and Chang, A.C.Y. (1977) *J. Bact.* 132, 724-737.
2. Cohen, S.N., Chang, A.C.Y., Boyer, H.W. and Helling, R.B. (1973) *Proc. Natl. Acad. Sci. USA* 70, 3240-3244.
3. Hasunuma, K. and Sekiguchi, M. (1977) *Mol. Gen. Genet.* 154, 225-230.
4. Fischhoff, D.A., Vovis, G.F. and Zinder, N.D. (1980) *J. Mol. Biol.* 144, 247-265.
5. Ohtsubo, H., Zenilman, M. and Ohtsubo, E. (1980) *J. Bact.* 144, 131-141.
6. Bernardi, A. and Bernardi, F. (1981) *Nucleic Acids Research* 9, 2905-2911.
7. Bernardi, F. and Bernardi, A. (1984) *Mol. Gen. Genet.* 195, 452-458.
8. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980) *J. Mol. Biol.* 143, 161-178.
9. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3963-3965.
10. Dente, L., Cesareni, G. and Cortese, R. (1983) *Nucleic Acids Research* 11, 1645-1655.
11. Birnboim, H.C. and Doly, J. (1979) *Nucleic Acids Research* 7, 1513-1523.
12. Wallace, R.B., Johnson, M.J., Suggs, S.V., Miyoshi, K., Bhatt, R. and Itakura, K. (1981) *Gene* 16, 21-26.
13. Maxam, A.M. and Gilbert, W. (1980) in *Methods in Enzymology*, Grossman, L. and Moldave, K. Eds., Vol. 65 part 1, pp. 499-560.
14. Joyce, C.M. and Grindley, N.D.F. (1983) *Proc. Natl. Acad. Sci.* 80, 1830-1834.
15. Greene, P.J., Heyneker, H.I., Bolivar, F., Rodriguez, R.I., Betlach, M.C., Covarrubias, A.A., Backman, K., Russel, D.J., Tait, R. and Boyer, H.W. (1978) *Nucleic Acids Research* 5, 2372-2380.
16. Kiss, A., Sain, B., Csordas-Toth, E. and Venetianer, P. (1977) *Gene* 1, 323-329.
17. Sutcliffe, G. (1978) *Cold Spring Harbor Symposia* 43, 77-90.
18. Peden, K.W.C. (1983) *Gene* 22, 277-280.
19. Polisky, B., Greene, P.J., Garfin, D.E., McCarthy, B.J., Goodman, H.M. and Boyer, H.W. (1975) *Proc. Natl. Acad. Sci. USA* 72, 3310-3314.
20. Bolivar, F., Rodriguez, R., Betlach, M.C. and Boyer, H.W. (1977) *Gene* 2, 75-93.
21. Miller, C.A., Tucker, W.T., Meacock, P.A., Gustafsson, P. and Cohen, S.N. (1983) *Gene* 24, 309-315.
22. Vocke, C. and Bastia, D. (1983) *Proc. Natl. Acad. Sci. USA* 80, 6557-6561.
23. Churchward, G., Linder, P. and Caro, L. (1983) *Nucleic Acids Research* 11, 5645-5659.
24. Armstrong, K.A., Acosta, R., Ledner, E., Machida, Y., Pancotto, M., McCormick, Ohtsubo, H., Ohtsubo, E. (1984) *J. Mol. Biol.* 175, 331-347.

25. Van Embden, J. and Cohen, S.N. (1973) *J. Bact.* 116, 699-709.
26. Crisona, N.J. and Clark, A.J. (1977) *Science* 196, 186-187.
27. Warren, G.J., Twigg, A.J. and Sherratt, D.J. (1978) *Nature* 274, 259-261.
28. Dougan, G., Saul, M., Warren, G. and Sherratt, D. (1978) *Molec. Gen. Genet.* 158, 325-327.
29. Nordheim, A., Hashimoto-Gotoh, T. and Timmis, K.N. (1980) *J. Bact.* 144, 923-932.
30. Guiney, D.G. and Yakobson, E. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3595-3598.
31. Finnegan, J. and Sherratt, D. (1982) *Molec. Gen. Genet.* 185, 344-351.