Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 6557–6561, November 1983
Biochemistry

Primary structure of the essential replicon of the plasmid pSC101

(nucleotide sequence analysis/*in vitro* mutagenesis/gene fusion)

CATHY VOCKE AND DEEPAK BASTIA*

Department of Microbiology and Immunology, Duke University Medical Center, Durham, NC 27710

Communicated by W. K. Joklik, August 8, 1983

ABSTRACT **The replicon of the low copy number plasmid pSC101 has an obligatory requirement for the *dnaA* initiator protein of *Escherichia coli* as well as a plasmid-encoded initiator protein. We have identified the cistron of the plasmid-encoded initiator by DNA sequence analysis. Fusion of the initiator cistron with the *lacZ* gene of *E. coli* yielded a fusion protein of ≈150 kilodaltons, thus confirming that the open reading frame detected by DNA sequence analysis actually encoded a 37.5-kilodalton protein. Deletion of 26 amino acid residues from the COOH terminus of the plasmid initiator abolished autonomous replication from pSC101 origin. By *in vitro* deletion analysis we have shown that, although sequences downstream from the initiator cistron are dispensable, a maximum of 400 base pairs immediately upstream from the NH_2-terminal region of the initiator is necessary for plasmid replication. These upstream sequences contain an A+T-rich region and three tandem repeats of a 21-base pair sequence; these features are characteristics of other replication origins.**

The plasmid pSC101, discovered by Cohen and co-workers (1, 2) is 9.4 kilobases in length, carries resistance to tetracycline, and replicates unidirectionally from a fixed origin (3). The replication of the plasmid, unlike that of Col E_1 (4), does not require DNA polymerase I (3). Furthermore, the replication initiation of the plasmid is mediated by the *dnaA* protein of *Escherichia coli* (5) and probably by a plasmid-encoded protein (6). The involvement of a plasmid-encoded initiator protein is suggested not only by the isolation of temperature-sensitive replication-defective mutants of the plasmid replicon (6) but also by *in vitro* replication studies. It is possible to replicate the plasmid *in vitro* in a cell-free extract that requires coupled transcription and translation of the plasmid DNA (J. Kaguni, R. Fuller, and A. Kornberg, cited in ref. 7; unpublished data). The dependence on both *dnaA* and a plasmid-encoded initiator protein makes the pSC101 replicon an interesting system to study from the point of view of initiation of replication. Previous work had identified the upper limits of the minimal replicon of pSC101 in a *Hae* II fragment that is located adjacent to the partition locus of the plasmid (8).

To analyze the initiation of replication of pSC101, we have further dissected the basic replicon by nucleotide sequence analysis and *in vitro* deletion analysis. The results reported in this paper show that an open reading frame that corresponds to the plasmid-encoded initiator protein cistron and sequences located upstream from the start of the open reading frame are essential for a functional replicon. Sequences located downstream from the open reading frame can be deleted without abolishing the ability of the remainder of the DNA to replicate autonomously. Deletion of parts of the open reading frame corresponding to the COOH-terminal amino acids of the putative plasmid-encoded protein abolishes autonomous replication.

MATERIALS AND METHODS

Bacterial, Phage, and Plasmid Strains. The *E. coli* strains JM103 (Δ *lac*, *pro* 1, *sup* E, *thi*, *end* A, *sbc* B15, *hsd* R4, *lac* I^q Z M15, F′*pro*A^+B^+) and MC1000 (*ara* D139, Δ (*ara*, *leu*), Δ *lac* Y74, *gal* U, *gal* k, *Str*A) were obtained from J. Messing and M. Casadaban, respectively. The plasmid pTU1, which contained the largest *Hin*cII fragment of pSC101 in the vector pACYC 184, was a gift from Stanley Cohen.

The plasmids pUC7, pUC8, and pUC9 were gifts from J. Messing as were the phage strains M13 mp7, mp8, mp9, mp10, and mp11 (9).

DNA Sequence Analysis. DNA sequence analysis was carried out by the method of Sanger *et al.* (10) with slight modifications as published (11). The secondary structure of the putative protein was predicted by the method of Chou and Fasman (12).

***In Vitro* Deletion Analysis.** The plasmid pCV2 that contained the *Hae* II replicator of pSC101 in between a *Hin*dIII and an *Eco*RI site was linearized by digestion with either *Eco*RI or *Hin*dIII and then digested under controlled conditions with either BAL-31 nuclease or with exonuclease III followed by digestion by nuclease S1. The shortened insert was then cut out with either *Hin*dIII or *Eco*RI digestion, purified by preparative gel electrophoresis, and ligated into either M13 mp8 or M13 mp9 replicative form DNAs that were doubly digested with *Hin*cII and *Eco*RI or with *Hin*cII and *Hin*dIII. Single-stranded DNAs from recombinant clones were prepared. The end points of the deletions were checked by performing G-specific dideoxy sequence reactions of the recombinant template DNAs by using a pentadeca-nucleotide universal primer.

An internal deletion, *Sau*Δ1 (see Fig. 2), was obtained by making a partial heteroduplex of single-stranded circular DNA of the *Hae* II replicator in M13 mp9 and appropriate restriction fragments belonging to the *Hae* II replicator, followed by digestion of the double-stranded region with *Sau*3A. The resected *Sau*3A-cut partial heteroduplex was then ligated to recircularize the DNA and transformed into *E. coli* JM103 (13).

Fusion with *lacZ*. The sequence corresponding to the putative replication protein was trimmed at the appropriate end with BAL-31 and then ligated to *Hin*dIII- and *Hin*cII-cut M13 mp11. Pale blue plaques on indicator plates were picked and the point of fusion on the DNA was determined by G-specific sequence reaction. The inserts were cut out from the recombinant M13 clones with *Bam*HI and *Hin*dIII and then transferred into the *lacZ* fusion vectors pORF5 and pMLB 1034 (14). The *Bam*HI site of M13 mp11 is such that any insert in the phage vector that is in the right reading frame with *lacZ*, when cut out with *Bam*HI and *Hin*dIII, also remains in the correct transla-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: bp, base pair(s); kDa, kilodalton(s); NH_2PhSGal-Sepharose, *p*-aminophenyl-β-D-thio-galactoside-Sepharose.

* To whom reprint requests should be addressed.

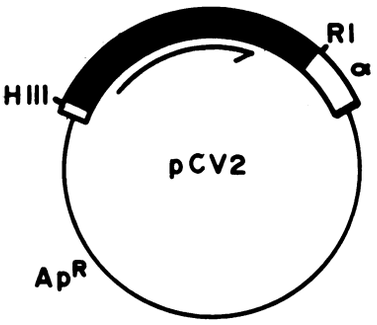

FIG. 1. The structure of the plasmid pCV2 that contains the *Hae* II replicator of pSC101 between the unique *Eco*RI and *Hin*dIII sites. α, α-Donor polypeptide cistron of β-galactosidase: ——, pUC9 sequence; ■, pSC101 sequences. The arrow indicates the open reading frame that corresponds to the plasmid initiator protein. Ap^R, β-lactamase gene. H III, *Hin*dIII; RI, *Eco*RI.

tional reading frame of *lacZ* in the unique *Bam*HI site of pORF5 and pMLB 1034 vectors.

Purification of β-Galactosidase Fusion Proteins. Purification was carried out as described (ref. 15; unpublished data).

Test for Autonomous Replication. The ability of a DNA fragment to replicate autonomously was checked by cloning the fragment into pUC9 and then examining the ability of the recombinant DNA to transform a *pol* A^- *E. coli* host. The DNA fragments were also ligated to a 1,500-base pair (bp) fragment that carried the gene for resistance to kanamycin but could not replicate by itself. The ligated DNA was transformed into *E. coli pol* A^+. The recovery of kanamycin-resistant transformants showed that the DNA fragment under examination could replicate autonomously.

RESULTS

Strategy for DNA Sequence Analysis. Previous work had established that an ≈1,800-bp-long *Hae* II fragment of pSC101 was capable of autonomous replication and that this segment lacked a functional partition locus (8). We cloned the 1,800-bp-long *Hae* II replicator fragment that was isolated by restricting the plasmid pTU1, which contained the largest *Hin*cII fragment of pSC101, into the vector pUC9 (9). The resulting recombinant clone pCV2 (Fig. 1) was capable of replication in a *pol*A^- mutant of *E. coli*, thus confirming that the cloned *Hae* II insert was capable of autonomous replication. The self-replicating insert was cut out with *Eco*RI and *Hin*dIII and cloned in both orientations in M13 mp8 and M13 mp9 vectors (9). The nucleotide sequence analysis of the *Hae* II fragment was carried out by the dideoxy chain-terminator method (10) by using template DNAs derived from the M13 recombinant clones and the various subclones. The sequencing strategy is shown in Fig. 2.

Nucleotide Sequence of the Minimal Replicon. The nucleotide sequence of the minimal replicon of pSC101 is shown in Fig. 3. The sequence is characterized by the presence of an open reading frame between the coordinates 503 and 1,450. The predicted amino acid sequence of the putative protein corresponding to this open reading frame is shown in Fig. 3. The secondary structure of this putative replication protein as predicted by the Chou–Fasman procedure (12) is shown in Fig. 4. Examination of the predicted structure shows several regions with α-helical structures, with the major α-helical areas located between the amino acids 104–178 and 260–310. The small stretches of predicted β sheets and turns are also shown. The predicted molecular mass of the protein is 37.4 kilodaltons (kDa), provided that the protein has the NH_2-terminus shown in Fig. 3. The putative replication protein has a net charge of +6, indicating that it should be weakly basic.

The analysis of codon usage in the putative replication protein shows frequent usage of rare codons. For example, percentage of synonym use of the codons UUA (leu), AUA (ser), ACA (thr), CGA (arg), AGA (arg), and AGG (arg) is 30%, 23%, 37%, 28%, 14%, 29%, and 43%, respectively, in the initiator but only 6%, 2%, 7%, 6%, 6%, 2%, 1%, and 0.25%, respectively, in other cistrons of *E. coli* (16). It has been suggested (16) that the more frequent usage of rare codons may be a mechanism by which the level of protein in the cell is kept low at the translational step.

The nucleotide sequence of the pSC101 replicon upstream from the open reading frame has three tandem repeats of a 21-bp sequence that are underlined in Fig. 3. Near the three repeats is a region that is 87% A+T-rich and is shown by a dotted line in Fig. 3.

There are several regions of striking, potential 2-fold symmetry that are located between the coordinates 436 and 514, 1,361 and 1,435, and 1,816 and 1,844 (Fig. 5). The largest of the three potential hairpins contains the putative initiator codon of the open reading frame.

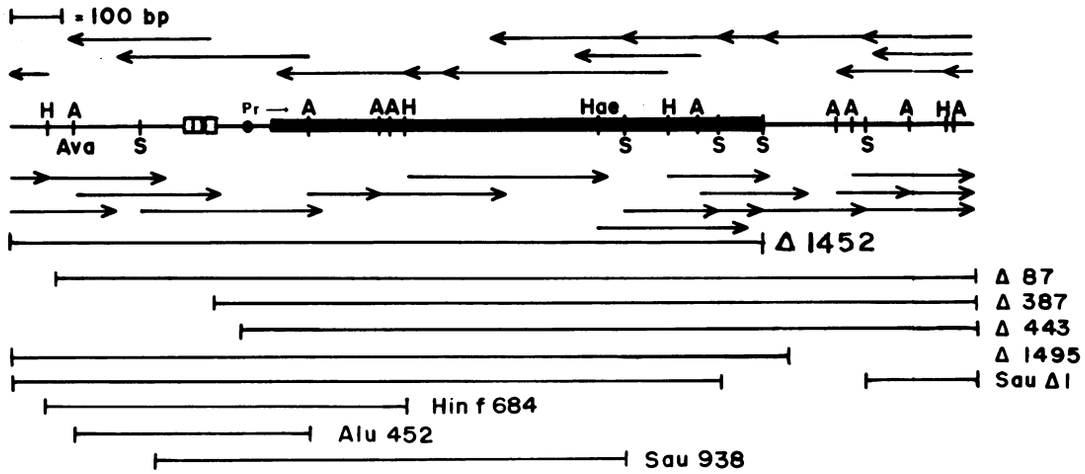

FIG. 2. Strategy used to sequence the replicon of pSC101. The arrows indicate the direction and extent of each sequence analysis reaction. The heavy line indicates the open reading frame. The three tandem repeats are indicated by the three boxes. Pr, putative promoter of the initiator cistron; H, *Hin*fI; A, *Alu* I; S, *Sau*3A; Ava, *Ava* I; Hae, *Hae* III. The locations of the various deletions are shown by bracketed lines. The lines indicate the sequence left intact in the deletions.

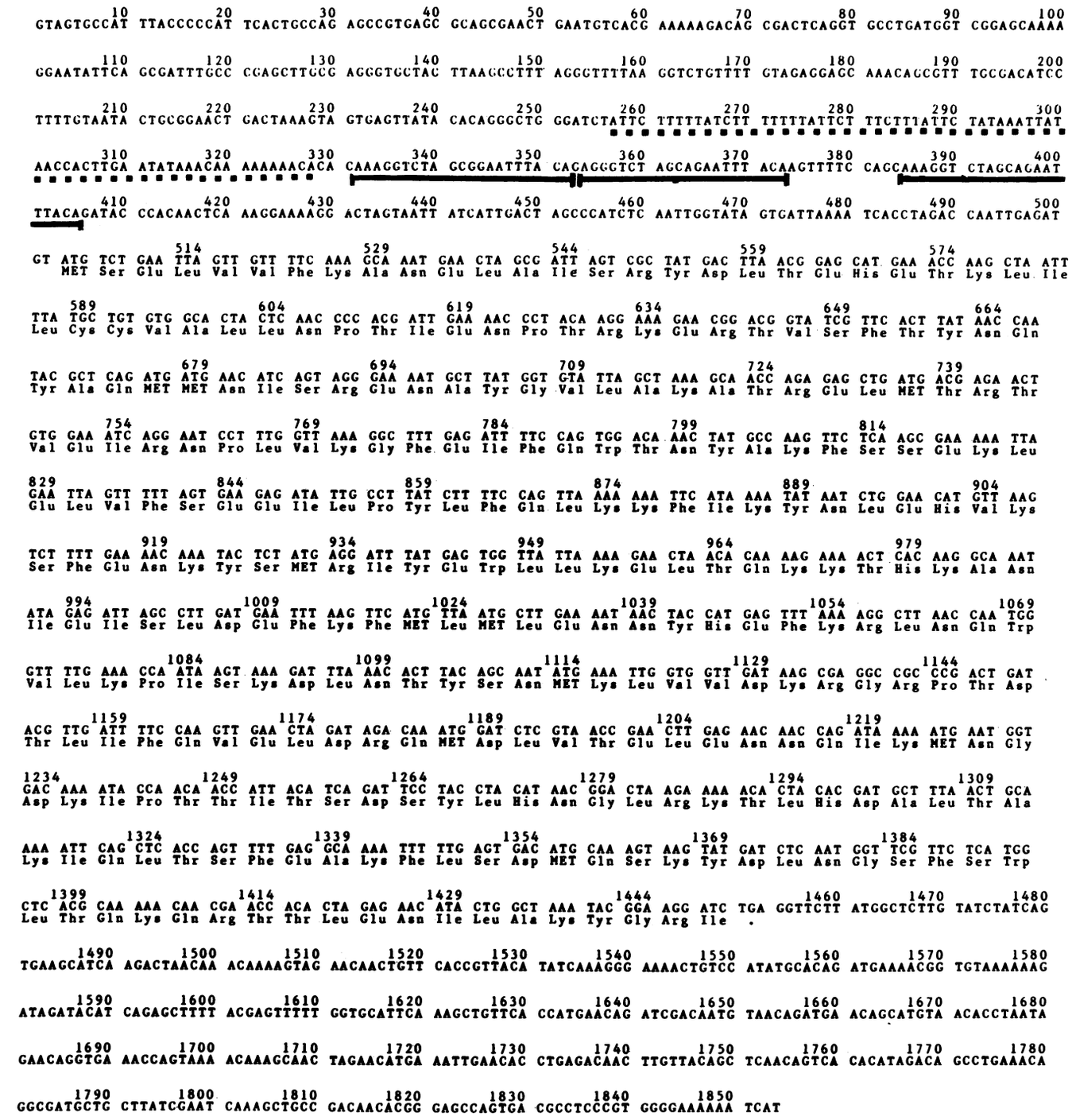

FIG. 3. The nucleotide sequence of the autonomously replicating region of pSC101. The predicted amino acid sequence of the plasmid-encoded initiator is shown. The dotted lines show the region that is 87% rich in A·T base pairs. The three sequences underlined by solid continuous lines are three tandem repeats, each 21 bp long.

***In Vitro* Deletion Analysis of the *ori* Region.** We have introduced several deletions extending various distances from either end of the *ori* sequence (Fig. 3). We have also introduced an internal deletion of the COOH-terminal region of the putative replication initiator cistron. These deletions are shown in Fig. 2. The results show that up to 87 bp from the coordinate 1 and the region immediately downstream from the end of the open reading frame (positions 1,453–1,854) are apparently dispensable, as shown by the fact that the remainder of the DNA fragment, after ligation to a nonself-replicating fragment that carried resistance to kanamycin, can transform *E. coli* cells to resistance to 25 μg of kanamycin per ml.

In contrast to the aforementioned regions, deletion of the sequence from the coordinates 87–387 abolishes replication. This deletion (Δ 387 in Fig. 2) removes the A+T-rich sequence and two of the three direct repeats described above. Similarly, the deleted subclones Δ 443, Hinf 684, Alu 452, and Sau 938 (Fig. 2) were all replication negative. In each of these cases we were unable to rescue the mutants by providing the putative plasmid-encoded replication protein in *trans*. SauΔ1, which re-

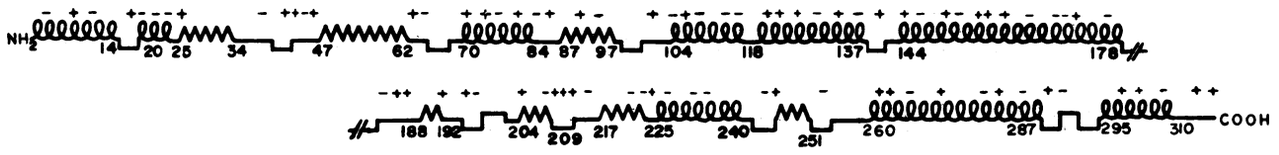

FIG. 4. The predicted secondary structure of the pSC101 initiator. , α helix; , β sheets; , turn. The numbers refer to the amino acid residues.

moved the segment of DNA from positions 1,370 to 1,643, was also replication negative, even in the presence of a helper plasmid.

In contrast to the mutants mentioned above, the mutant Δ 1,495, which lacked the sequence between the coordinates 1,495–1,854, was capable of autonomous replication, as was the mutant Δ 1,452, which had lost all of the sequence downstream from the open reading frame. From the aforementioned deletion analysis, we conclude that the upper limit of the minimal essential replicon of pSC101 probably consists of the sequences from the coordinates 87–1,450, which includes the A+T-rich region, the three direct repeats, the promoter region of the putative replication protein, and the reading frame of the putative protein. All other sequences are apparently dispensable. Curiously, none of the replication mutants could be rescued by a helper plasmid by providing the plasmid-encoded replication protein in *trans*.

The Open Reading Frame Codes for a Protein. To examine if the open reading actually encoded a protein, we fused the open reading frame at the coordinate 1,452 in the correct translational frame to the *lacZ* cistron located in the plasmid pORF5 (14). The point of the gene fusion had the following sequence:

	initiator protein		β-galactosidase	
1,445				
AGG	ATC	TGG	ACT	CTA
Arg	Ile	Trp	Thr	Leu

We prepared, from cells carrying the *lacZ* fusion plasmid,

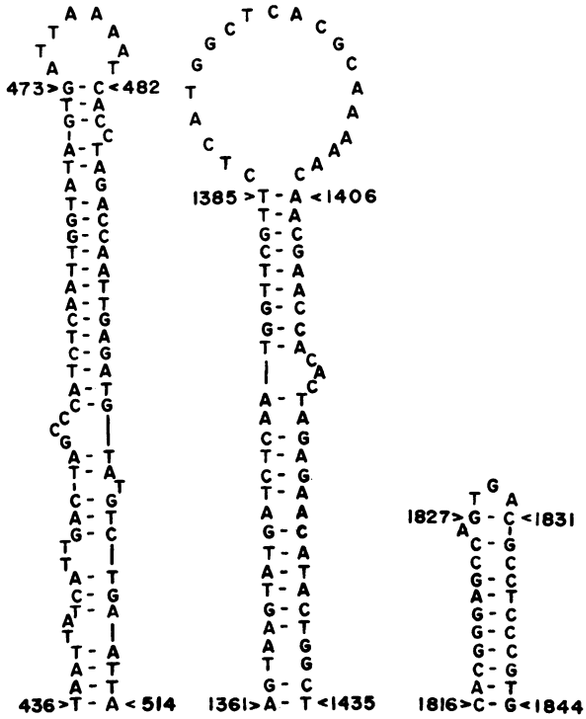

FIG. 5. The potential hairpin sequence of the *ori* regions of pSC101.

the β-galactosidase-tagged, plasmid-encoded replication protein by affinity chromatography through *p*-aminophenyl-β-D-thio-galactoside-Sepharose columns (NH_2PhSGal-Sepharose) and analyzed the protein fractions by NaDodSO_4/polyacrylamide gel electrophoresis. The results depicted in Fig. 6 show that the affinity column enriches a protein of ≈150 kDa from protein fractions of cells that carry the *lacZ*-pSC101 fusion plasmid (Fig. 6). In contrast, the 150-kDa protein band is not visible in protein fractions that were prepared from cells that did not carry the *lacZ* fusion plasmid. On the basis of the enrichment of the 150-kDa protein band from cells that carry the *lacZ* fusion plasmid by β-galactosidase-specific affinity column chromatography, we conclude that the open reading frame from the coordinates 503–1,450 does indeed code for a protein. The protein is estimated to be ≈37 kDa, which would be consistent with that predicted from the nucleotide sequence of the open reading frame.

DISCUSSION

On the basis of the nucleotide sequence analysis of an autonomously replicating fragment of pSC101 DNA and resection of the replicator by *in vitro* deletion analysis we have determined that the functional replicon consists of a maximum of 400 bp of

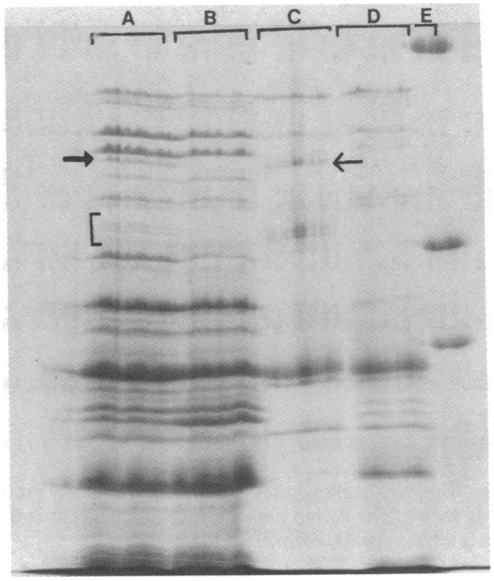

FIG. 6. NaDodSO_4/polyacrylamide gel profiles of proteins. Lanes: A, 40% $(NH_4)_2SO_4$ protein fraction of the *lacZ* fusion plasmid containing *E. coli*; B, 40% $(NH_4)_2SO_4$ protein fraction from *E. coli* that did not harbor the *lacZ* fusion plasmid; C, *lacZ* fusion protein after NH_2PhSGal-Sepharose chromatography; D, protein fraction described for lane B after the NH_2PhSGal-Sepharose step; E, markers: myosin (200 kDa), β-galactosidase (116 kDa), phosphorylase B (92.5 kDa). The arrow indicates the pSC101 initiator–*lacZ* fusion protein, which is ≈150 kDa. [, Possible degradation products of the fusion protein. The fusion protein and its probable degradation products are enriched after the NH_2PhSGal-Sepharose purification step.

DNA followed by an open reading frame that has the potential to code for a 37.5-kDa protein.

The sequences upstream from the start of the open reading frame are characterized by an A+T-rich segment followed by three direct repeats of 21 bp each in length. A+T-rich regions and direct repeats have also been found in the replication origins of phage λ (17, 18), R6K (19), RK2 (20), F factor (21), etc. In the case of λ (22) and R6K (23, 24), the replication initiator protein binds to the directly repeated sequences. Furthermore, in the case of F factor, the direct repeat sequences are involved in plasmid incompatibility (21, 25).

The open reading frame that is potentially capable of coding for a 37.5-kDa protein is essential for replication, as shown by the fact that deletions that removed base pairs equivalent to 26 or more codons from the region corresponding to the putative COOH terminus abolished autonomous replication. However, the replication-defective mutants could not be rescued by a helper plasmid that should have provided the normal 37.5-kDa protein in *trans*. Therefore, it is not possible at this time to distinguish between the alternatives that the region of the open reading frame corresponding to the COOH terminus of the putative 37.5-kDa initiator protein is necessary as a critical region of a *trans*-acting protein or as a *cis*-acting site essential for pSC101 replication. The discovery of the open reading frame in the replicator region of pSC101 suggested that the plasmid initiator protein might be encoded by the open reading frame. A more definitive proof that the open reading frame actually encoded a 37.5-kDa protein is provided by the generation of a 150-kDa fusion protein by plasmids that contain the entire open reading frame mentioned above fused in the correct translational phase with the *lacZ* cistron of *E. coli*. This fusion protein was readily detected on $NaDodSO_4$/polyacrylamide gels after the protein fractions were chromatographed in the β-galactosidase-specific $NH_2PhSGal$-Sepharose affinity column. Recently we have discovered that the fusion protein specifically binds to the region of pSC101 replication origin located upstream from the start of the 37.5-kDa replication protein cistron (unpublished data).

In addition to the plasmid-encoded initiator protein, pSC101 has an obligatory requirement for the *dnaA* initiator protein of *E. coli* (5). It remains to be seen whether *dnaA* protein binds to the replication origin of pSC101. We have compared the nucleotide sequence of the *ori* region reported here with that of *oriC* of *E. coli* (26, 27) and found the following direct repeats that are in common between the two origins: the sequences T-T-A-T-A-C-A-C-A located between the coordinates 236 and 244 and T-T-A-T-A-A-C-C-A between 297 and 305 (Fig. 3) are similar to the sequences T-T-A-T-C-C-A-C-A, T-T-A-C-A-C-A, T-T-A-T-C-C-A-A-A, and T-T-A-T-C-C-A-C-A that are found at the coordinates 80–88, 186–194, 229–237, and 260–268 of the *oriC* sequence reported by Hirota *et al.* (26). It remains to be seen as to what extent this homology is biologically important.

Meacock and Cohen (8) have reported that a short region preceding the sequence of the minimal origin of pSC101 controls the partition of the plasmid. This partition locus (*par*) has recently been subjected to complete sequence analysis by Cohen and co-workers (28). A very short region of the *par* sequence overlaps the first 120 bp of the *ori* sequence reported here. There does not seem to be any striking homology between the *par* sequence and the sequences that are essential for a functional replicon of pSC101.

We are grateful to Dr. S. N. Cohen and his co-workers for sharing with us strains and for communicating to us the *par* locus sequence prior to its publication. We thank Mr. Duncan McRee for his help with the secondary structure prediction and Mrs. S. Shipman for typing the manuscript. This work was supported by Grant 1PO1-CA-30246-02 awarded by the National Cancer Institute. D.B. is an Established Investigator of the American Heart Association.

1. Cohen, S. N. & Chang, A. C. Y. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 1293–1297.
2. Cohen, S. N. & Chang, A. C. Y. (1977) *J. Bacteriol.* **132**, 734–737.
3. Cabello, F., Timmis, K. & Cohen, S. N. (1976) *Nature (London)* **259**, 285–290.
4. Kingsbury, D. & Helinski, D. (1970) *Biochem. Biophys. Res. Commun.* **41**, 1538–1544.
5. Hasanuma, D. & Sekiguchi, M. (1977) *Mol. Gen. Genet.* **154**, 225–230.
6. Hashimoto, T. & Sekiguchi, M. (1976) *J. Bacteriol.* **127**, 1561–1563.
7. Kornberg, A. (1983) *Supplement to DNA Replication* (Freeman, San Francisco).
8. Meacock, P. & Cohen, S. N. (1980) *Cell* **20**, 529–542.
9. Vieira, J. & Messing, J. (1982) *Gene* **19**, 259–268.
10. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
11. Germino, J. & Bastia, D. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 5475–5479.
12. Chou, P. & Fasman, G. (1978) *Adv. Enzymol.* **47**, 45–148.
13. Muller, U. & Wells, R. D. (1980) *J. Mol. Biol.* **141**, 1–24.
14. Schultz, J., Silhavy, T. J., Berman, M. L., Fiil, N. & Emr, S. (1982) *Cell* **31**, 227–235.
15. Steers, E., Cuatrecasas, P. & Anfinsen, C. (1971) *J. Biol. Chem.* **246**, 196–200.
16. Konigsberg, W. & Godson, N. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 687–691.
17. Moore, D., Denniston-Thompson, K., Kruger, K., Furth, M., Williams, B., Daniels, D. & Blattner, F. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **43**, 155–163.
18. Hobom, G., Grosschedl, R., Lusky, M., Scherer, G., Schwartz, F. & Kossel, H. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **43**, 165–178.
19. Stalker, D. M., Kolter, R. & Helinski, D. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1150–1154.
20. Stalker, D., Thomas, C. & Helinski, D. (1982) *Mol. Gen. Genet.* **181**, 8–13.
21. Murotsu, T., Matsubara, K., Sugisaki, H. & Takanami, M. (1981) *Gene* **15**, 257–271.
22. Tsurimoto, T. & Matsubara, K. (1981) *Nucleic Acids Res.* **9**, 1789–1799.
23. Germino, J. & Bastia, D. (1983) *Cell* **32**, 131–140.
24. Germino, J. & Bastia, D. (1983) *Cell* **34**, 125–134.
25. Tolun, A. & Helinski, D. (1981) *Cell* **24**, 687–694.
26. Hirota, Y., Yasuda, M., Nishimura, A., Sugimoto, K., Sugisaki, H., Oka, A. & Takanami, M. (1981) in *Initiation of DNA Replication*, ICN–UCLA Symposium on Molecular and Cellular Biology, ed. Ray, D. (Academic, New York), Vol. 22, pp. 1–11.
27. Messer, W., Meijer, M., Bergmans, H. E. N., Hansen, F. G., VonMeyenberg, K., Beck, E. & Schaller, H. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **43**, 139–145.
28. Miller, C. A., Tucker, W. T., Meacock, P. T., Gustafsson, P. & Cohen, S. N. (1983) *Gene* **24**, 309–315.