Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 5475–5479, September 1982
Biochemistry

Primary structure of the replication initiation protein of plasmid R6K

(gene cloning/recombinant DNA/dideoxy sequence analysis/symmetric replication)

JOSEPH GERMINO AND DEEPAK BASTIA*

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

Communicated by Irwin Fridovich, May 25, 1982

ABSTRACT The cistron of the replication initiation protein of plasmid R6K has been cloned into the single-strand DNA vectors M13mp8 and M13mp9 and its complete nucleotide sequence has been determined. The amino acid sequence of the initiator protein as predicted from its nucleotide sequence shows that the protein is lysine rich and weakly basic and has a molecular weight of 35,000, which is in close agreement with that estimated from the mobility in NaDodSO$_4$/acrylamide gels. The secondary structure of the protein, approximated by the probabilistic methods of Chou and Fasman [Chou, P. & Fasman, G. (1978) *Adv. Enzymol.* 47, 45–148], suggests an NH$_2$-terminal domain of primarily positively charged α-helical structure, a core region of interspersed short stretches of random coils and β-sheets and -turns, and a COOH-terminal domain of α-helix.

The mechanism of the initiation of replication of double-stranded DNA molecules that replicate in a topologically symmetric configuration (i.e., both strands of the DNA are replicated by similar or identical mechanisms), from a specific origin is a major unsolved problem in molecular biology. A class of proteins called initiator proteins are involved in the potentiation of the first replication forks at the origin. A detailed understanding of the structure, enzymatic properties, and site of action (on the DNA) of this class of proteins is essential for unraveling the molecular mechanism of initiation of replication.

The drug-resistance factor R6K, which confers resistance to ampicillin and streptomycin and has a molecular weight of 26 million (1), replicates in a Cairns type configuration from multiple origins of replication (2, 3). An *in vitro* system for replicating R6K DNA and its derivatives is available (4, 5); the plasmid encodes its own initiator protein (5, 6) and contains a specific replication terminus (7, 8). These characteristics recommend the plasmid as a convenient system to study the mechanisms of initiation and termination of replication.

The initiation protein of R6K has been shown to be encoded in the DNA sequence contained in the *Hin*dIII fragments 9 and 15 (5, 6). Subcloning experiments had established that a combination of the restriction fragments *Hin*dIII 9-15-2 and 15-9-4 yielded self-replicating miniplasmids (5) having the replication origins α and γ (2, 3) located in the *Hin*dIII fragment 4 and the replication origin β located in the *Hin*dIII fragment 2 (refs. 2, 6; unpublished work).

To study the molecular biology of the initiation of replication in R6K, we have cloned the cistron of this protein in the single-strand phage vectors M13mp5, M13mp8, and M13mp9, and in this report we present the complete nucleotide sequence of the cistron and the amino acid sequence of the initiator protein as predicted from its nucleotide sequence.

MATERIALS AND METHODS

Bacterial Strains, Phage Strains, and Plasmids. *Escherichia coli* strain JM 103 (*Δlac, pro-1, SupE, thi, endA, sbcB15, hsdR4, lacI^q, lacZm13,* F′ *proA$^+$, proB$^+$*) and the M13 strains mp8 and mp9 were obtained from J. Messing through the Bethesda Research Laboratory. The recombinant plasmid pJG3 contains the *Hin*dIII fragments 2, 15, and 9 of R6K cloned into the *Hin*dIII site of pBR322.

Enzymes. T4 DNA ligase and T7 gene *6* exonuclease were purified according to published procedures (9, 10). The restriction endonuclease *Alu* I was purified as described (11). Restriction endonucleases *Bam*HI, *Bgl* II, and *Hin*dIII were purchased from New England BioLabs and Bethesda Research Laboratories. *Hae* II was a gift from Cathy Vocke. The Klenow fragment of *E. coli* DNA polymerase I was purchased from Boehringer.

Biochemicals. Most of the standard biochemicals were purchased from Sigma. [α-^{32}P]dATP for DNA sequence analysis was purchased from Amersham (400 Ci/mmol; 1 Ci = 3.7 × 10^{10} becquerels). The pentadecanucleotide universal primer and dideoxynucleotide triphosphates were purchased from New England BioLabs and P-L Biochemicals, respectively.

DNA Sequence Analysis. Both strands of the DNA were analyzed by the method of Sanger (12) using single-stranded DNA templates of clones of the various restriction fragments in M13mp8 or M13mp9 vectors. A part of the sequence was also determined by using a template generated by T7 exonuclease (10). The DNA sequence data were analyzed by the Molgen (Stanford University)–SUMEX AIM computer facility (National Institutes of Health).

The secondary structure of the protein was predicted by the probabilistic method of Chou and Fasman (13).

RESULTS AND DISCUSSION

Strategy for Nucleotide Sequence Analysis. Molecular cloning and complementation experiments had previously shown that (*i*) all minireplicons of R6K had restriction fragments *Hin*dIII 9 and 15 in common (ref. 6 and Fig. 1) and (*ii*) although the replication origins α and β could not be made to initiate DNA replication by providing a diffusible gene product (i.e., initiator protein) in *trans*, the replication origin γ of R6K could readily be complemented by the provision of the gene product from the *Hin*dIII 9–15 region cloned into a second plasmid vector (5). These observations strongly suggested that the structural gene for the initiator protein was located, at least partly, in the *Hin*dIII 9–15 region of the R6K chromosome.

We attempted to locate more precisely the cistron for the initiator protein by nucleotide sequence analysis. The plasmid

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.

* To whom reprint requests should be addressed.

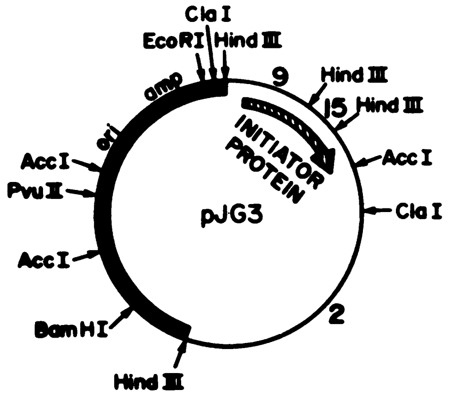

FIG. 1. Physical structure of the plasmid chimera pJG3. The plasmid contains the *Hin*dIII fragments 9, 15, and 2, which constitute the β replicon of R6K cloned at the *Hin*dIII site of the vector pBR322. The plasmid chimera can replicate in a pol Ats host at the restrictive temperature. ■, pBR322; —, R6K.

pJG3 (Fig. 1) was used to prepare preparative amounts of the restriction fragments needed. All sequence analyses were carried out by using the chain-terminator method of Sanger (12). The procedure requires specific primers and single-stranded DNA templates. The single-stranded templates were prepared by digesting linear pJG3 DNA (which was linearized at either the unique *Eco*RI site or the BamHI site) with gene 6 exonuclease of phage T7 (8). Alternatively, the restriction fragments from the region to be analyzed were cloned into M13mp5, mp8, and mp9 vectors and recombinant phage particles were used as the source of single-stranded DNA templates (14). In the case of exonuclease-generated templates, restriction fragments that were less than 150 base pairs long were used as internal primers for the DNA analysis reactions. All M13 recombinant single-stranded templates were analyzed by using a synthetic pentadecamer as the universal primer (14). The strategy for DNA sequence analysis is summarized in Fig. 2.

During our attempts to clone the region shown in Fig. 2 into M13 vectors, we noticed that any restriction fragment that contained the putative promoter region (coordinates 8–240) could be cloned into the M13 vectors in only one orientation. For example, although the *Hin*dIII fragment 15 could be cloned into the M13mp5 vector in both orientations of the insert, the *Hin*dIII fragment 9 could be recovered in only one orientation in the recombinant clones. Furthermore, when the *Hin*dIII fragment 9 was subdivided by digestion with *Alu* I and the subfragments were cloned into the *Hin*cII site of M13mp8 and mp9, all subfragments except the promoter-containing subfragment (coordinates 8–240 in Figs. 2 and 3) could be cloned in both orientations. Apparently the recombinants containing one specific orientation of the putative promoter region of the initiator protein of R6K are lethal, presumably due to interference in M13 replication.

Both strands of the DNA corresponding to all regions of the initiator protein cistron (Fig. 3), except the first 200 base pairs, were analyzed by the procedures mentioned above. The sequence of the first 200 base pairs was derived by inspection of at least four separate sequence gels with an unambiguous pattern of bands, albeit from one strand of the DNA. The accuracy of the DNA sequence was checked by confirming the presence of additional restriction enzyme recognition sites, predicted from the sequence, by gel electrophoresis of DNA restricted with the appropriate endonucleases.

Characteristics of the DNA Sequence. The sequence of the putative cistron region for the initiator protein of R6K is shown in Fig. 3. The DNA sequence is relatively A+T rich and the longest open reading frame of the sequence starts with the ATG codon located at 181 and ends at the TGA codon at 1096. All other possible reading frames are interrupted by multiple chain terminators at various points and are therefore considerably shorter. The longest open reading frame identified above predicts a protein having a molecular weight of 35,000, which is consistent with the size of the initiator protein as estimated by $NaDodSO_4$ gel electrophoresis (ref. 6; unpublished results).

The nucleotide sequence contains several regions of dyad symmetry, which are listed in Fig. 3. Two of the dyad symmetries are located between the coordinates 65 and 149 (Fig. 3), which is in the untranslated leader region of the initiator protein locus.

We have examined the frequency of codon usage of the replication initiator protein cistron from the predicted amino acid sequence and compared it with those of the ribosomal protein (15), of the lacI (16), lacY (17), trpA (18), and recA (19, 20) proteins, and the lipoprotein cistrons of *E. coli* (21). It has been suggested that the frequency of codon usage in *E. coli* ribosomal protein cistrons reflects the frequencies of occurrence of the major species of tRNA synthetases; more frequent use of those codons corresponding to the major species of the tRNA synthetase was suggested to facilitate efficient translation (15). However, comparison of the codon usage of the replication initiator with that of the ribosomal protein cistron (15) and the other *E. coli* proteins mentioned above shows certain departures from this rule. For example, we found that the codon CGU is frequently used by the other *E. coli* proteins but not used at all by the initiator protein. Conversely, the codon UUA is very frequently used by the initiator but infrequently used by the other proteins.

In the sequence shown in Fig. 3, there are three ATG codons

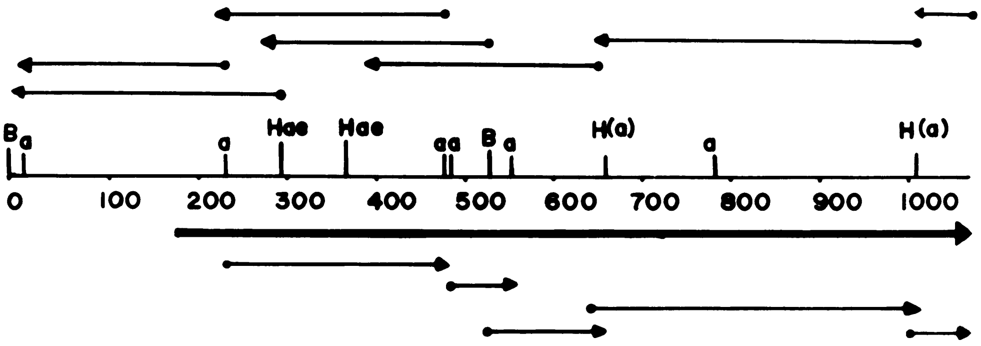

FIG. 2. Strategy for DNA sequence analysis. Arrows indicate extent and direction of sequences obtained in separate experiments. The heavy arrow marks the coding region of the initiator protein. Numbers refer to nucleotide pairs. Hae, *Hae* II; B, *Bgl* II; a, *Alu* I; H, *Hin*dIII sites; *Hin*dIII sites are (left to right) fragments 9, 15, and 2.

```
        10         20         30         40         50         60         70
AGATCTAGCT TAAAACAGGT GGCTTTTTAA TCATCTTTGC CAAGCATGGC GCGGGTTTGG GGTAATATAG

        80         90        100        110        120        130        140
CGACTCATAA AAGCGTTAAA CATGAGTGGA TAGTACGTTG CTAAAACATG AGATAAAAAT TGACTCTCAT

       150        160        170        180                  195
GATATTGGCG TTAAGATATA CAGAATGATG AGGTTTTTTT  ATG AGA CTC AAG GTC ATG ATG
                                             MET Arg Leu Lys Val MET MET

        210                 225                 240                 255
GAC GTG AAC AAA AAA ACC AAA ATT CGC CAC CGA AAC GAG CTA AAT CAC ACC CTG
Asp Val Asn Lys Lys Thr Lys Ile Arg His Arg Asn Glu Leu Asn His Thr Leu

                270                 285                 300
GCT CAA CTT CCT TTG CCC GCA AAG CGA GTG ATG TAT ATG GCG CTT GCT CCC ATT
Ala Gln Leu Pro Leu Pro Ala Lys Arg Val MET Tyr MET Ala Leu Ala Pro Ile

    315                 330                 345                 360
GAT AGC AAA GAA CCT CTT GAA CGA GGG CGA GTT TTC AAA ATT AGG GCT GAA GAC
Asp Ser Lys Glu Pro Leu Glu Arg Gly Arg Val Phe Lys Ile Arg Ala Glu Asp

            375                 390                 405
CTT GCA GCG CTC GCC AAA ATC ACC CCA TCG CTT GCT TAT CGA CAA TTA AAA GAG
Leu Ala Ala Leu Ala Lys Ile Thr Pro Ser Leu Ala Tyr Arg Gln Leu Lys Glu

420                 435                 450                 465
GGT GGT AAA TTA CTT GGT GCC AGC AAA ATT TCG CTA AGA GGG GAT GAT ATC ATT
Gly Gly Lys Leu Leu Gly Ala Ser Lys Ile Ser Leu Arg Gly Asp Asp Ile Ile

        480                 495                 510                 525
GCT TTA GCT AAA GAG CTT AAC CTG CCC TTT ACT GCT AAA AAC TCC CCT GAA GAG
Ala Leu Ala Lys Glu Leu Asn Leu Pro Phe Thr Ala Lys Asn Ser Pro Glu Glu

                540                 555                 570
TTA GAT CTT AAC ATT ATT GAG TGG ATA GCT TAT TCA AAT GAT GAA GGA TAC TTG
Leu Asp Leu Asn Ile Ile Glu Trp Ile Ala Tyr Ser Asn Asp Glu Gly Tyr Leu

    585                 600                 615                 630
TCT TTA AAA TTC ACC AGA ACC ATA GAA CCA TAT ATC TCT AGC CTT ATT GGG AAA
Ser Leu Lys Phe Thr Arg Thr Ile Glu Pro Tyr Ile Ser Ser Leu Ile Gly Lys

            645                 660                 675
AAA AAT AAA TTC ACA ACG CAA TTG TTA ACG GCA AGC TTA CGC TTA AGT AGC CAG
Lys Asn Lys Phe Thr Thr Gln Leu Leu Thr Ala Ser Leu Arg Leu Ser Ser Gln

690                 705                 720                 735
TAT TCA TCT TCT CTT TAT CAA CTT ATC AGG AAG CAT TAC TCT AAT TTT AAG AAG
Tyr Ser Ser Ser Leu Tyr Gln Leu Ile Arg Lys His Tyr Ser Asn Phe Lys Lys

        750                 765                 780                 795
AAA AAT TAT TTT ATT ATT TCC GTT GAT GAG TTA AAG GAA GAG TTA ATA GCT TAT
Lys Asn Tyr Phe Ile Ile Ser Val Asp Glu Leu Lys Glu Glu Leu Ile Ala Tyr

                810                 825                 840
ACT TTT GAT AAA GAT GGA AAT ATT GAG TAC AAA TAC CCT GAC TTT CCT ATT TTT
Thr Phe Asp Lys Asp Gly Asn Ile Glu Tyr Lys Tyr Pro Asp Phe Pro Ile Phe

    855                 870                 885                 900
AAA AGG GAT GTG TTA AAT AAA GCC ATT GCT GAA ATT AAA AAG AAA ACA GAA ATA
Lys Arg Asp Val Leu Asn Lys Ala Ile Ala Glu Ile Lys Lys Lys Thr Glu Ile

            915                 930                 945
TCG TTT GTT GGC TTC ACT GTT CAT GAA AAA GAA GGA AGA AAA ATT AGT AAG CTG
Ser Phe Val Gly Phe Thr Val His Glu Lys Glu Gly Arg Lys Ile Ser Asn Leu

960                 975                 990                1005
AAG TTC GAA TTT GTC GTT GAT GAA GAT GAA TTT TCT GGC GAT AAA GAT GAT GAA
Lys Phe Glu Phe Val Val Asp Glu Asp Glu Phe Ser Gly Asp Lys Asp Asp Glu

       1020                1035                1050                1065
GCT TTT TTT ATG AAT TTA TCT GAA GCT GAT GCA GCT TTT CTC AAG GTA TTT GAT
Ala Phe Phe MET Asn Leu Ser Glu Ala Asp Ala Ala Phe Leu Lys Val Phe Asp

               1080                1095
GAA ACC GTA CCT CCC AAA AAA GCT AAG GGG TGA
Glu Thr Val Pro Pro Lys Lys Ala Lys Gly  .
```

FIG. 3. Nucleotide sequence of the coding region and the noncoding leader region of the initiator protein. The predicted amino acid sequence is shown. The putative ribosome binding site is underlined. Regions having dyad symmetries are as follows: 91–104/77–65, 135–149/123–109, 274–286/271–258, 364–375/361–351, 959–972/956–941, 960–975/954–939, 1034–1051/1015–997, 1073–1089/1059–1044, where the numbers in the numerator refer to coordinates of the sequence having the ability to form base pairs with the sequence specified by the numbers in the denominator.

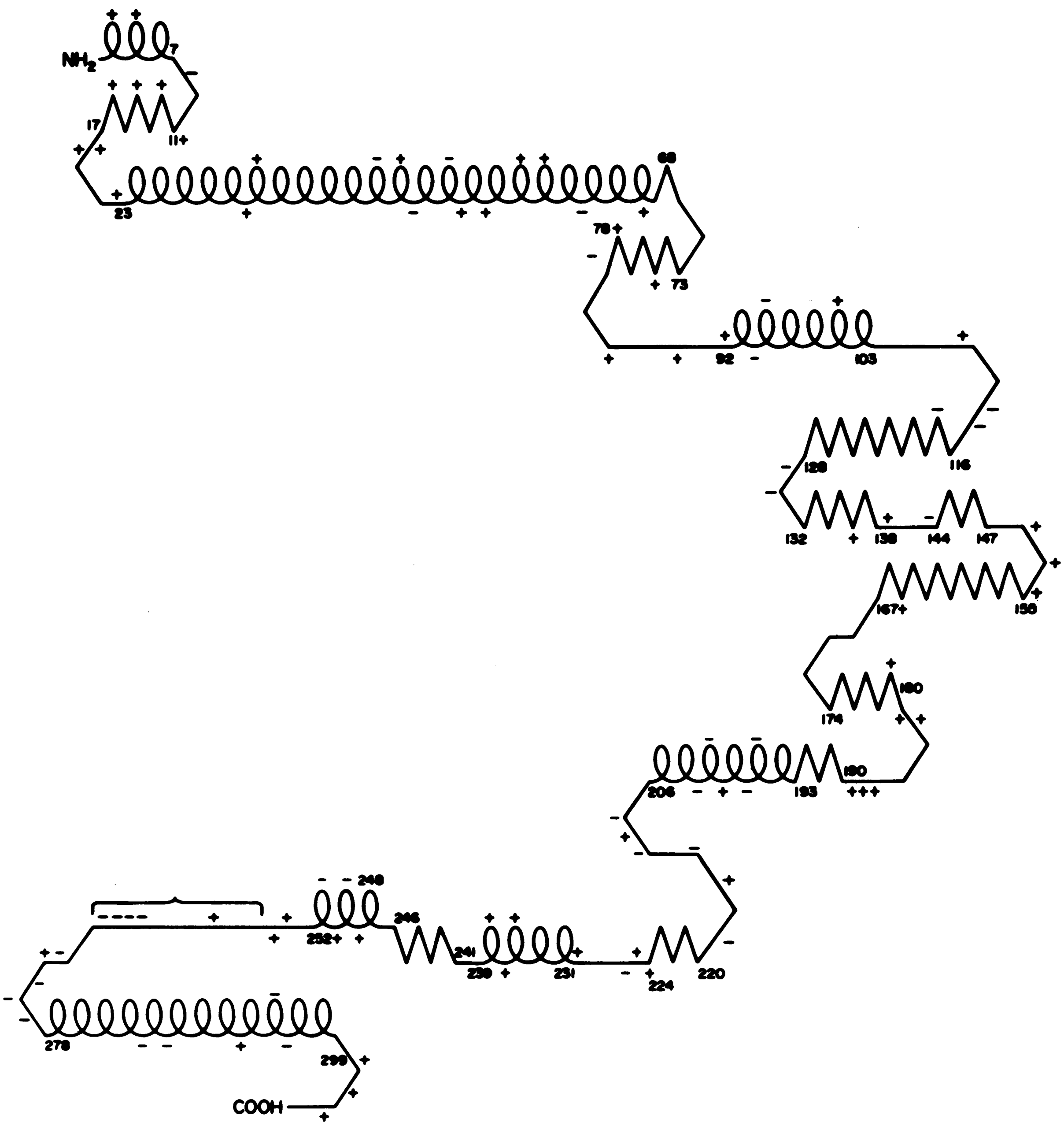

FIG. 4. Predicted secondary structure of the replication initiation protein. ꝏ, α-helix; /\/\, β-sheet; ——, random coil,), turn. The region near the COOH-terminal end, marked by a large brace, has a structure that could not be predicted with any reasonable confidence.

very close to the putative NH_2-terminal end of the initiator protein. To determine which of the three ATG codons is the real initiation codon, NH_2-terminal amino acid sequence analysis of the initiator protein will be necessary. Construction of overproducer strains in which the relevant cistron is linked to an efficient promoter and ribosome binding site would greatly facilitate purification of the initiator protein.

Fusion of the NH_2-Terminal Region of the Initiator to the COOH-Terminal Segment of β-Galactosidase. To determine the correct reading frame of the initiator protein by an independent method, we attempted to fuse the initiator protein cistron with the *lacZ* gene contained in the M13mp7 vector. *Hin*dIII fragment 9 was cleaved at the *Bgl* II sites located at coordinates 1 and 528 (Figs. 2 and 3) and cloned by ligation into the *Bam*HI site of the M13mp7 vector. The fragment was recovered in only one orientation in the recombinant clone, which had the following sequence:

```
                                        lacZ
              525
. . . . . . CCT GAA GAG TTA GAT CCG TCG . . . . . . .
   .   .  Pro - Glu - Glu - Leu - Asp - Pro - Ser   .   .
            R6K initiator
```

The recombinant, which should have the NH_2-terminal segment of the initiator protein fused to the COOH-terminal segment of β-galactosidase in the correct reading frame, does indeed produce a hybrid protein with β-galactosidase activity. This is shown by the fact that the recombinant clones produce blue plaques on 5-bromo-4-chloroindolyl β-galactoside plates in the presence of isopropyl thiogalactoside. Thus, the gene fusion experiment confirms the correct reading frame of the cistron deduced from the nucleotide sequence data.

The Predicted Amino Acid Sequence. The predicted amino acid sequence of the initiator protein is shown in Fig. 3. The protein contains 55 positively charged and 44 negatively charged residues. Therefore, the protein should be weakly basic. A striking feature of the protein is the large number of lysine residues, which constitute approximately 12% of the total amino acids.

The predicted secondary structure of the protein derived according to the probabilistic methods of Chou and Fasman (13) is shown in Fig. 4. The protein is predicted to have 38% α-helices, 22% β-sheets and 40% random coils. With the understanding that the predictive method is 70% correct, the following features of the protein warrant discussion.

The NH_2-terminal domain of the protein contains the longest helical region having a net positive charge. This region appears to be similar to the NH_2-terminal region of another initiator protein, namely the O protein of phage λ (22). The NH_2-terminal region of the O protein appears to recognize a specific nucleotide sequence at the replication origin (23). It is tempting to predict that the NH_2-terminal region of the initiator protein of R6K may have a similar function.

In addition to the feature mentioned above, the initiator protein of R6K appears to contain a core region of primarily β-sheets mixed with random coils, turns, and short α-helical regions. The COOH-terminal region appears to be a negatively charged α-helix. This region apparently is dispensable because recombinant DNA clones containing the *Hin*dIII 9–15 fragments of R6K lack the COOH-terminal region yet produce a functional initiator protein (6). It is reasonable to assume that the initiator protein, besides recognizing specific nucleotide sequences of the origin regions, interacts with other proteins of the replisome. The domain for the protein–protein interaction may reside in the β-sheeted random-coiled core region of the protein.

Attempts to purify the protein by conventional methods were frustrated by the low copy number of the protein per cell and its apparent instability. Overproduction of the protein by genetic engineering should facilitate its purification and detailed study of its exact role in the molecular mechanism of initiation of replication at the replication origins of R6K. The availability of the complete nucleotide sequence of the initiator cistron should facilitate not only attempts to overproduce the protein but also experiments to study the functions of its predicted domains by site-directed mutagenesis.

We thank Mrs. Elizabeth Eubanks for typing the manuscript and Mr. D. McRee for help with the secondary structure calculations. This work was supported by U.S. Public Health Service Grant 1PO1-CA30246-01. J.G. was supported by Medical Scientist Training Program Grant 5T32GM07171 from the U.S. Public Health Service.

1. Kontomichalou, P., Mitani, M. & Clowes, R. C. (1970) *J. Bacteriol.* **104**, 34–44.
2. Crosa, J. H., Luttropp, L., Heffron, F. & Falkow, S. (1975) *Mol. Gen. Genet.* **140**, 39–50.
3. Inuzuka, N., Inuzuka, M. & Helinski, D. R. (1980) *J. Biol. Chem.* **255**, 11071–11074.
4. Inuzuka, M. & Helinski, D. R. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 5381–5385.
5. Kolter, R., Inuzuka, M., Figurski, D., Thomas, C., Stalker, D. & Helinski, D. R. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **43**, 91–103.
6. Crosa, J. H., Luttropp, L. R. & Falkow, S. (1978) *J. Mol. Biol.* **124**, 443–468.
7. Germino, J. & Bastia, D. (1981) *Cell* **23**, 681–687.
8. Bastia, D., Germino, J., Crosa, J. & Ram, J. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2095–2099.
9. Murray, N. E., Bruce, S. A. & Murray, K. (1979) *J. Mol. Biol.* **132**, 493–505.
10. Kerr, C. & Sadowski, P. D. (1972) *J. Biol. Chem.* **247**, 305–310.
11. Roberts, R. J., Meyer, P. A., Morrison, A. & Murray, K. (1976) *J. Mol. Biol.* **102**, 157–165.
12. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
13. Chou, P. & Fasman, G. (1978) *Adv. Enzymol.* **47**, 45–148.
14. Messing, J., Crea, R. & Seeburg, P. H. (1981) *Nucleic Acids Res.* **9**, 309–321.
15. Post, L. E. & Nomura, M. (1980) *J. Biol. Chem.* **255**, 4660–4666.
16. Farabaugh, P. J. (1978) *Nature (London)* **274**, 765–769.
17. Buchel, D. F., Gronenberg, B. & Muller-Hill, B. (1980) *Nature (London)* **283**, 541–545.
18. Nichols, B. P. & Yanofsky, C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5244–5248.
19. Horii, T., Ogawa, T. & Ogawa, H. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 313–317.
20. Sancar, A., Stachelek, C., Konigsberg, W. & Rupp, W. D. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2611–2615.
21. Nakamura, K., Pirtle, R. M., Pirtle, I. C., Takeishi, K. & Inouye, M. (1980) *J. Biol. Chem.* **255**, 210–216.
22. Scherer, G. (1978) *Nucleic Acids Res.* **5**, 3141–3156.
23. Furth, M. E., McLeester, C. & Dove, W. F. (1978) *J. Mol. Biol.* **126**, 227–240.