Proc. Natl. Acad. Sci. USA
Vol. 75, No. 3, pp. 1081–1085, March 1978
Biochemistry

Nucleotide sequences of the separate origins of synthesis of bacteriophage G4 viral and complementary DNA strands

(DNA replication/RNA priming/chain termination sequencing)

JOHN C. FIDDES*, B. G. BARRELL, AND G. NIGEL GODSON†‡

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Communicated by F. Sanger, November 21, 1977

ABSTRACT **Bacteriophage G4 has physically separated origins of synthesis of its viral and complementary DNA strands. Chain termination and "plus and minus" DNA sequencing methods have been used to obtain the nucleotide sequence of these two origins. The unique origin at which the complementary DNA strand is initiated has been located in the untranslated region between genes *F* and *G*. This sequence, which has considerable secondary structure, contains a stretch which is complementary to the RNA primer that is observed during synthesis *in vitro* of the G4 complementary DNA strand [Bouché, J. P., Rowen, L. & Kornberg, A. (1978) *J. Biol. Chem.*, in press]. This G4 origin shows extensive sequence homology with the bacteriophage λ origin of DNA replication [Denniston-Thompson, K., Moore, D. D., Kruger, K. E., Furth, M. E. & Blattner, F. R. (1977) *Science* 198, 1051–1056]. The sequence around the site in gene *A* at which G4 viral DNA strand synthesis is initiated by the nicking action of the cistron *A* protein is very similar to that of bacteriophage φX174. An (A+T)-rich stretch flanked by (G+C)-rich sequences may be involved in the interaction between the DNA and protein.**

G4 is an isometric φX-like phage isolated in 1974 (1). Its single-stranded DNA genome is approximately 5580 nucleotides long and codes for the same series of nine proteins as φX174 (1).

G4 has a unique origin of synthesis for its complementary DNA strand, both for the conversion of the single-stranded viral DNA to double-stranded replicative form (RF) *in vitro* (2) and during RF synthesis *in vivo* (3–5). This initiation process requires *dnaG* protein and DNA binding protein, both from *Escherichia coli* (2, 6, 7), and involves the synthesis of an RNA primer molecule (2), the nucleotide sequence of which has now been determined (8). In contrast, φX174 has multiple origins of synthesis of its complementary DNA strand and requires *dnaB, dnaC,D,* i, and n proteins from *E. coli,* in addition to *dnaG* protein and DNA binding proteins for its initiation (9, 10).

G4 also has a unique origin of synthesis of its viral DNA strand (11), and this is located at a different site from the origin of synthesis of the complementary DNA strand, approximately 100° away on the circular genome (4, 12, 13). On the aligned φX174/G4 restriction enzyme cleavage map,§ the origin of synthesis of the G4 viral DNA strand lines up with the equivalent origin of synthesis of the φX174 viral strand [which is in cistron *A* (14)] and the origin of the G4 complementary strand lines up with the φX174 untranslated region between genes *F* and *G* (15).

MATERIALS AND METHODS

Preparation of G4 DNA and Restriction Enzyme Fragments. G4 RF I DNA was prepared as described previously (16) and the strands were separated by the method of Greenfield *et al.* (17).

RF I DNA was digested with the following restriction enzymes: *Hae* III (*Haemophilus aegyptius*), a gift of A. Smith; *Hin*fI (*Haemophilus influenzae* serotype f), a gift of J. Ravetch and H. D. Robertson; *Taq* I (*Thermus aquaticus*), a gift of J. I. Harris and M. Runswick; *Hha* I (*Haemophilus haemolyticus*) and *Alu* I (*Arthrobacter luteus*), both purchased from BioLabs; and *Hin*dII (*Haemophilus influenzae* serotype d). DNA fragments were separated on polyacrylamide gradient gels (18) in TBE buffer (19). Fragments were eluted by electrophoresis (20).

DNA Sequence Analysis. The plus and minus procedure (21) was used as described by Barrell (22), and the chain termination method of DNA sequence analysis was that of Sanger *et al.* (23).

RESULTS AND DISCUSSION

Sequence of the G4 Origin of Viral DNA Strand Synthesis. The origin of G4 viral strand synthesis has been located in the restriction enzyme fragment *Hae* III 2a (4, 11–13). On the aligned G4 and φX174 restriction enzyme cleavage map (Fig. 1) G4 fragment *Hae* III 2a covers φX174 fragment *Hae* III 6b, which is known to contain the origin of φX174 viral DNA strand synthesis (14, 26, 27). The φX174 viral strand origin has recently been located precisely on the complete φX174 DNA sequence of Sanger *et al.* (25) by the observation that the φX174 cistron *A* protein cleaves the viral strand of φX174 RF I DNA between nucleotides 4297 (G) and 4298 (A) (28). The nucleotide sequence of the corresponding region of G4 DNA has now been determined using the chain termination method of Sanger *et al.* (23).

The appropriate restriction endonuclease fragments to use as primers were selected from an examination of the G4 restriction enzyme map (see Fig. 2 for the map of this region). The autoradiographs of the sequencing gels obtained when *Hha* I fragment 6 and *Alu* I fragment 4 were used as primers with viral strand DNA as template are shown in Fig. 3 *a* and *b*. These two priming sites are located downstream of the origin. Also shown (Fig. 3*c*) is the autoradiograph of a priming using a *Hae* III-generated subproduct of *Taq* I fragment 8 (*Taq* I ▽8) with complementary strand DNA as template. This priming site is located upstream of the origin.

With one region of ambiguity, a sequence 72 nucleotides long

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

Abbreviation: RF, replicative form.

* Present address: Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143.
† Permanent address: Department of Therapeutic Radiology, Yale University Medical School, 333 Cedar Street, New Haven, CT 06510.
‡ To whom reprint requests should be addressed, at Yale.

§ J. Grindley and G. N. Godson, unpublished data.

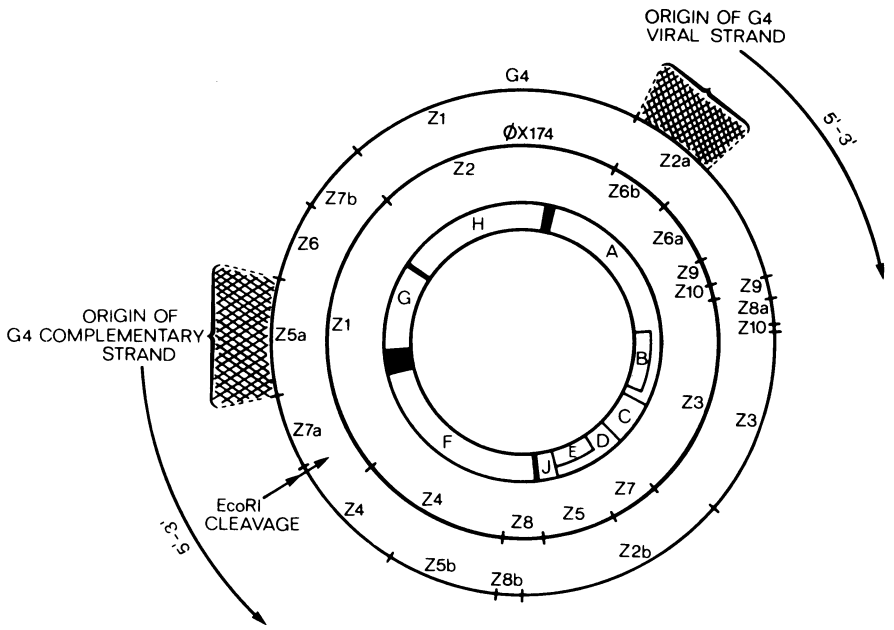

FIG 1. Location of the origin of G4 viral and complementary DNA strand synthesis. The G4 *Hae* III restriction cleavage map (outer circle) is from Godson (24), corrected from our unpublished sequencing data. The alignment of the G4 and ϕX174 restriction enzyme cleavage maps is from Grindley and Godson.[§] The ϕX174 restriction enzyme cleavage map (second circle) and genetic map (inner circle) are from Sanger *et al.* (25). The origin and direction of G4 complementary DNA strand synthesis are from the *in vitro* studies of Zechel *et al.* (2) and from the *in vivo* studies of Godson (3, 5) and Martin and Godson (4). The origin and direction of the G4 viral DNA strand synthesis are from the studies of Ray and Duebner (11, 12), Martin and Godson (4), and Godson (3, 13).

(8–79 in Fig. 4) can be interpreted from the *Taq* I ▽8 priming and is shown alongside the autoradiograph in Fig. 3*c*. This sequence was extended by seven nucleotides (1–7) towards the priming site in another experiment (not shown) in which the chain-terminated products were electrophoresed without being released by restriction enzyme cleavage from the 135-nucleotide-long primer molecule.

Nucleotide sequences 40 nucleotides (36–75) and 33 nucleotides (64–96) long, respectively, are shown in Fig. 3 *a* and *b* alongside the *Hha* I 6 and *Alu* I 4 autoradiographs. The *Alu* I 4 sequence overlaps completely with the *Taq* I ▽8 sequence and resolves the ambiguity. The *Alu* I 4 and *Hha* I 6 primings overlap each other by 12 nucleotides (64–75).

The nucleotide sequence of this region is shown in Fig. 4 and is displayed against the corresponding ϕX174 sequence. Because this sequence codes for the cistron *A* protein, the predicted amino acid sequence for both phages is also given.

A 30-nucleotide stretch of sequence around and including the cistron *A* cleavage site in ϕX174 (positions 4291–4320) is completely conserved in G4. This is one of the largest stretches of unchanged nucleotides in the G4 genome (unpublished data) and is located in an area of G4 DNA that is otherwise considerably different from ϕX174 (see Fig. 4). This suggests that the conserved sequence represents a protein recognition site that is common to the two phage origins of replication. It is not clear whether the protein is the viral cistron *A* protein or a host cell protein.

The DNA sequence of this region, even outside the fully conserved portion, preserves in both phages the arrangement of an (A+T)-rich (low melting temperature) region at the origin flanked by (G+C)-rich (high melting temperature) regions on either side (see Fig. 4). It is noticeable that the C_6,T pyrimidine tract (4285–4291) that was observed (29) to be present in the gap in the viral strand of late ϕX174 RF II molecules is absent in G4.

Sequence of the Origin of Synthesis of G4 Complementary DNA Strand. The *in vitro* origin of G4 complementary DNA strand synthesis is uniquely located approximately 5% of the genome from the single *Eco*RI cleavage site, on the opposite side of the genome from the viral DNA strand origin [(2) and see Fig. 1]. The *in vivo* origin of G4 complementary DNA strand synthesis appears to be in the same place and has been located in fragment *Hae* III 5a close to the *Hae* III 7a/5a junction (3–5). The nucleotide sequence of this region has been determined mainly by the chain termination method. The restriction enzyme fragment map of this region of G4 DNA is

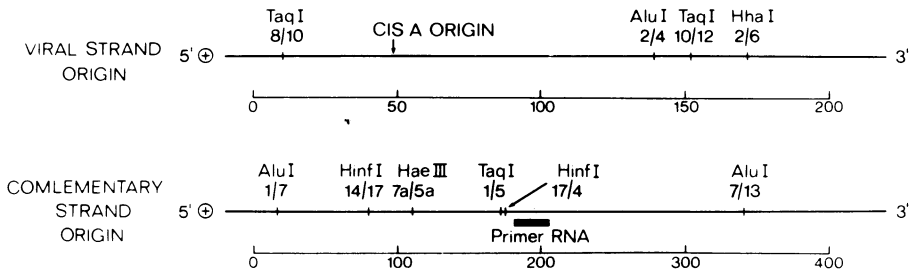

FIG. 2. Restriction enzyme cleavages round the G4 origin of viral and complementary DNA strand synthesis. These data were taken from Godson (24) and from the detailed restriction cleavage map derived from DNA sequencing studies (G. N. Godson and J. C. Fiddes, unpublished data). The two scales are in base pairs, with arbitrary zero points.

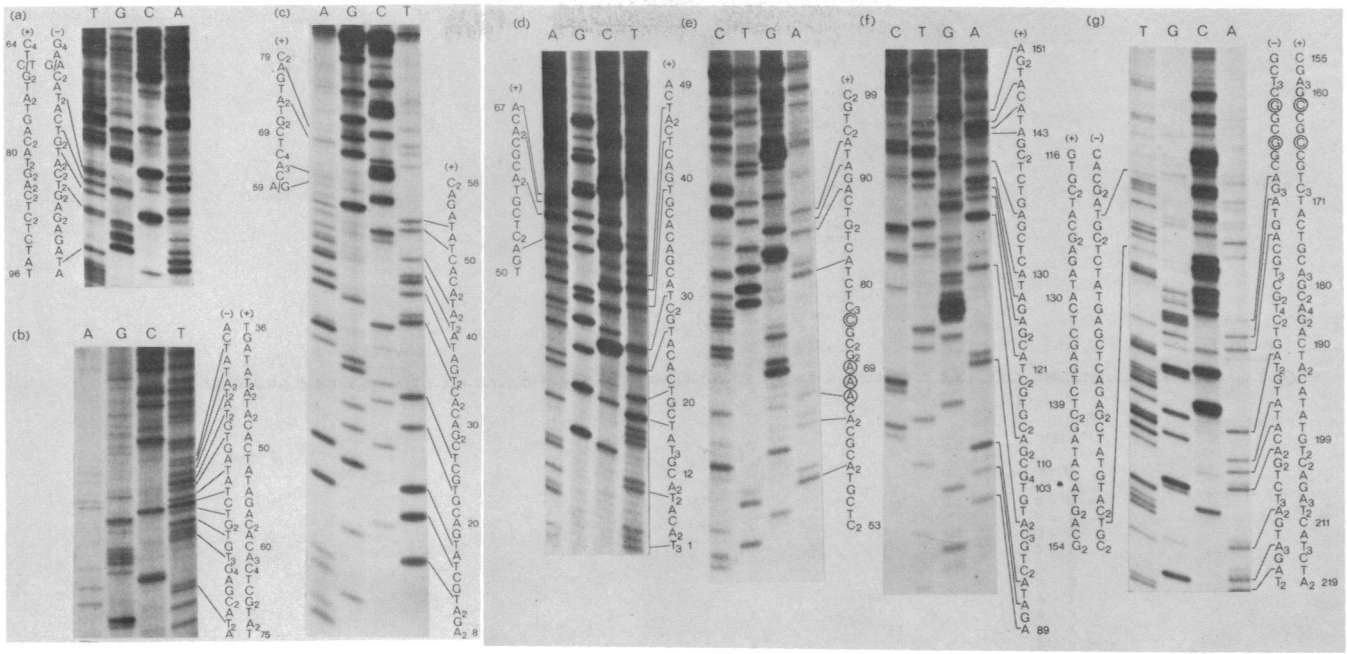

FIG. 3. Autoradiographs of DNA sequencing gels along with the deduced sequences. Ambiguities, which are discussed in the text, are shown by circling the relevant nucleotides. The chain termination method of Sanger *et al.* (23) was used in all cases. Chain terminating analogues were 2′,3′-dideoxy-ATP, 2′,3′-dideoxy-GTP, 2′,3′-dideoxy-TTP, and arabinosyl-CTP. (*a*) *Hha* I 6 viral strand, (*b*) *Alu* I 4 viral strand, (*c*) *Taq* I ∇8 complementary strand, (*d*) *Alu* I 1 complementary strand, (*e*) *Hinf* I 14 complementary strand, (*f*) *Hae* III ∇7a complementary strand, (*g*) *Alu* I 13 viral strand. Gels used were 8% (*a, b, d, g*) or 12% (*c, e, f*) polyacrylamide/7 M urea with TBE buffer (19). Electrophoresis was carried out at 45 mA to ensure that the gels became hot (60°–70°) in order to help maximize denaturation of the DNA. The numbering of the nucleotides is arbitrary.

shown in Fig. 2. Three fragments were used as primers with complementary strand DNA as template. These were *Alu* I 1, *Hin*fI 14 and a *Hin*dII subproduct of *Hae* III 7a (∇7a), all of which provided priming sites upstream of the origin. Fragment *Alu* I 13 was used with viral strand DNA as template to provide a priming site downstream of the origin. Examples of autoradiographs of these four primings are shown in Fig. 3 along with the deduced sequence.

A 67-nucleotide-long sequence can be interpreted from the *Alu* I 1 priming (Fig. 3*d*), and this overlaps by 15 nucleotides the *Hin*fI 14 priming (53–99) which is shown in Fig. 3*e*. However, two regions of the *Hin*fI 14 autoradiograph are hard to interpret (see Fig. 3*e*). The A-A-A (67–69) sequence preceding the *Hae* III 7a/5a site was verified by a *Hin*fI 14 plus and minus system priming (not shown) and by a *Hin*fI 4 priming using viral strand as template (not shown). The ambiguity immediately after the *Hae* III 7a/5a site is resolved by the absence of a *Hpa* II site in this sequence. Both of these problems are considered to be related to the secondary structure postulated in this region (see loop III in Fig. 5).

A 63-nucleotide-long sequence (89–151) that overlaps that of the *Hin*fI 14 priming by 11 nucleotides was obtained from the *Hae* III ∇7a priming (Fig. 3*f*). This sequence has a 36-base overlap with the 104 nucleotides (116–219) obtained from the *Alu* I 13 priming in the opposite direction (Fig. 3*g*). Over one region of the *Alu* I 13 autoradiograph the distance between successive bands becomes considerably reduced. This feature is considered to be due to the sequence of the newly synthesized DNA being such that it can form an intramolecular secondary structure loop causing anomalous electrophoretic migration. This compression of bands has been observed previously with the strongest hairpin loops in φX174 (15, 22, 31). The sequence of this region of G4, which corresponds to loop number 1 of Fig. 5 (149–169) was confirmed by a priming with *Hin*fI 17 using complementary strand as template (data not shown).

Most of this sequence corresponds to the untranslated region of the G4 genome, between genes *F* and *G* (see Fig. 5), and contained within the untranslated region is a sequence (147–174) complementary to the 28-nucleotide-long RNA primer molecule sequenced by Bouché *et al.* (8). This RNA primer is synthesized during *in vitro* complementary strand DNA synthesis (6). The 5′ end of the primer is at an A nucleotide (174)

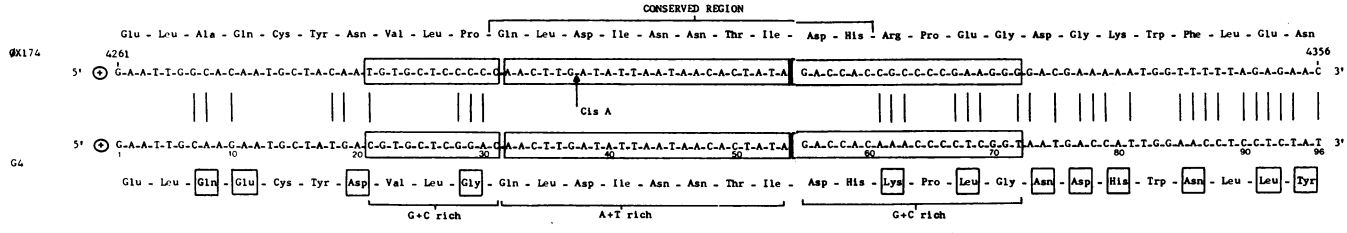

FIG. 4. Cistron *A* origin of G4 viral DNA strand synthesis. The G4 viral DNA strand sequence (5′-3′ transcription, left to right) was deduced as described in the text and the amino acid phasing was deduced from homology with φX174 cistron *A* protein. The viral strand φX174 nucleotide sequence and amino acid phasing are from Sanger *et al.* (25), as is the nucleotide numbering. The base sequence differences between G4 and φX174 are marked with lines and the changed amino acids are boxed. The site of the φX174 cistron *A* cleavage is from Langeveld *et al.* (28). The arbitrary numbering of the G4 nucleotides agrees with that in Fig. 3 *a*, *b*, and *c*.

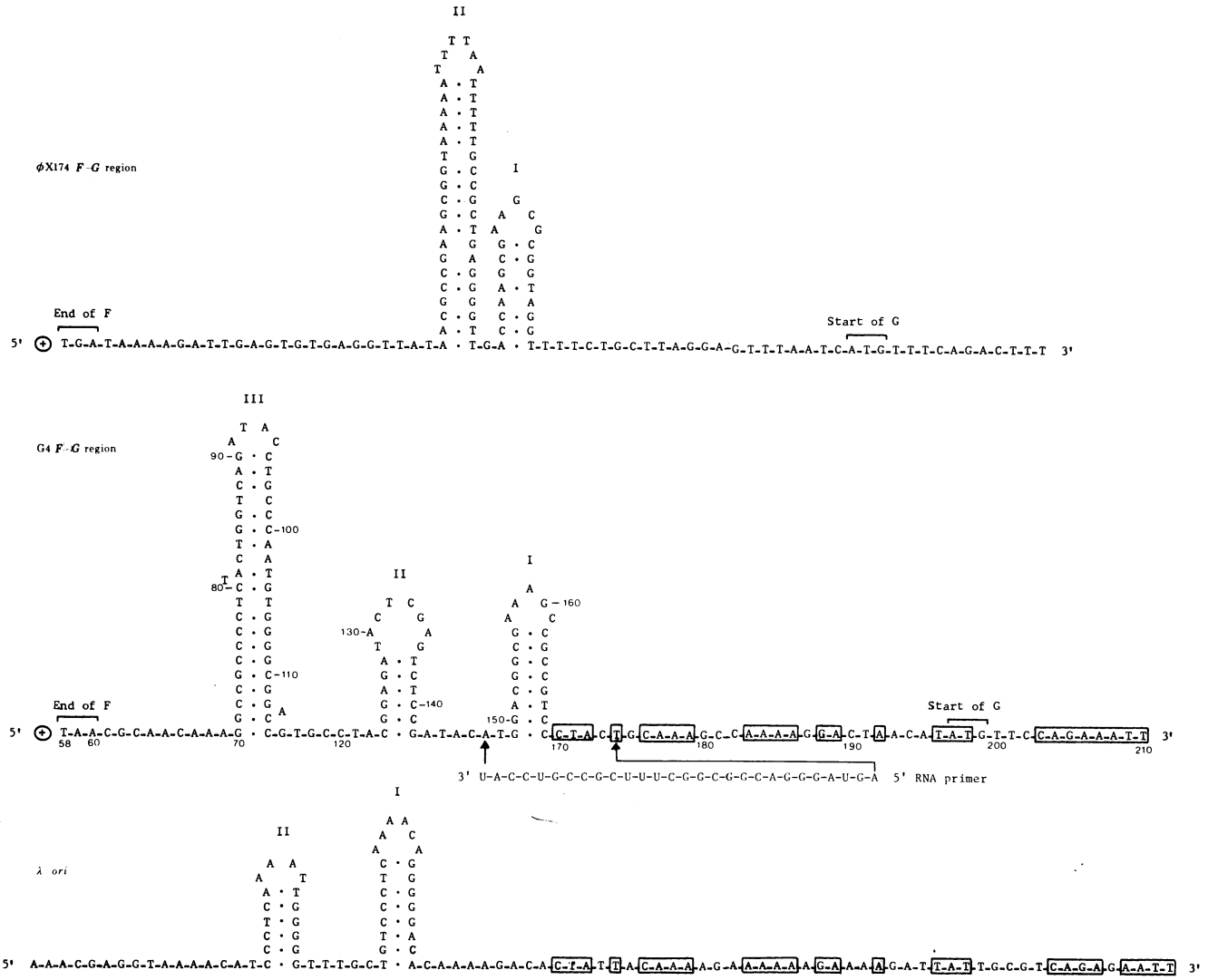

FIG. 5. Comparison of the proposed secondary structures of (*Upper*) φX174 *F–G* untranslated region (15), (*Middle*) G4 *F–G* untranslated region, which contains the RNA primer (sequence from ref. 8), and (*Lower*) a region of the origin of DNA replication in bacteriophage λ (*ori*) (30). The λ *ori* is shown with two hairpin loops instead of the single loop proposed by Denniston-Thompson *et al.* (30). The considerable sequence homology between the λ *ori* and the G4 sequence is shown. The sequence of the φX174 *F–G* region differs (F. Sanger, personal communication) from that previously published (15). The sequence C-G has been inserted between positions 2303 and 2304 (25).

23 bases from the A-T-G initiation codon of gene *G*, and this precisely defines the position of the complementary strand origin.

The G4 complementary strand origin does not have the same sequence features as the viral strand origin. The RNA primer is located in a (G+C)-rich region that can form a stable base-paired loop (loop I in Fig. 5). Two more stable secondary structure loops (loops II and III of Fig. 5) can be drawn in this region. The RNA primer molecules are heterogeneous in size (8), the longest ones ending at an A nucleotide before loop II. This loop may be the signal for the change to adding deoxyribonucleotide precursors, though deoxyribonucleotide precursors can be added at sites closer to the origin (8).

The structural differences observed between the origins of G4's viral and complementary DNA strand syntheses probably reflect different initiation processes. Initiation of DNA synthesis at the G4 viral strand origin requires specific nicking of the viral strand by the cistron *A* protein, followed by denaturation of the DNA by *rep* protein and DNA binding protein (32), while initiation of DNA synthesis at the complementary strand origin requires DNA binding protein and *dnaG* protein to synthesize an RNA primer.

Nucleotide Sequence Similarity between Bacteriophage λ Origin and the G4 Complementary DNA Strand Origin. It is interesting to note that while there are no obvious sequence similarities between the G4 origins and those of colicinogenic factor ColE1 (33) and bacteriophage f1 (34), both of which use RNA polymerase to prime DNA synthesis, there is considerable homology between the G4 complementary DNA strand origin and the bacteriophage λ origin (*ori*) (30), both of which use *dnaG* protein to initiate synthesis. This comparison is shown in Fig. 5, the region of homology being to the 3′ side of the G4 primer sequence from nucleotides 170 to 210. If the λ origin DNA is considered as single-stranded DNA (which may, in fact, be the case during the initiation of DNA synthesis), it can also form two stable base-paired loops similar to those of G4. The common DNA base sequences therefore probably represent the *dnaG* protein binding site. The hairpin loops may also function in the initiation process.

Comparison of the G4 and φX174 *F* and *G* Intercistronic

Regions. The untranslated region between genes *F* and *G* in G4 is 136 nucleotides long, compared with 110 nucleotides in ϕX174 (15). There is no base sequence homology or similarity, though both regions can be drawn with secondary structure (Fig. 5). The DNA sequences of the flanking structural genes F and *G*, however, show considerable similarity (unpublished data). The ϕX174 *F–G* intercistronic space contains eleven termination codons (in all phases), whereas the G4 space contains only one. ϕX174 gene *F* ends in a double termination signal, T-G-A-T-A-A, whereas G4 gene *F* ends in a single termination codon, T-A-A. Despite the difference in base sequence homology of the two regions, the A-G-G-A ribosome binding site is the same in both genomes and occurs seven nucleotides from the A-T-G start of gene *G*.

We thank Dr. F. Sanger for his advice and encouragement and for a gift of chain termination compounds. We also thank R. Staden for help with the computer handling and analysis of DNA sequences. G.N.G. was on sabbatical leave from the Department of Therapeutic Radiology, Yale Medical School, and was supported by United States Public Health Service Grant CA06519 and American Cancer Society Grant SG-86.

1. Godson, G. N. (1974) *Virology* **58**, 272–289.
2. Zechel, K., Bouché, J. P. & Kornberg, A. (1975) *J. Biol. Chem.* **250**, 4684–4689.
3. Godson, G. N. (1975) in *DNA Synthesis and Its Regulation*, eds. Goulian, M., Hanawalt, P. & Fox, C. F. (W. A. Benjamin, Menlo Park, CA), pp. 386–397.
4. Martin, D. M. & Godson, G. N. (1977) *J. Mol. Biol.* **117**, 321–336.
5. Godson, G. N. (1977) *J. Mol. Biol.* **117**, 337–352.
6. Bouché, J. P., Zechel, K. & Kornberg, A. (1975) *J. Biol. Chem.* **250**, 5995–6004.
7. Wickner, S. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 2815–2819.
8. Bouché, J. P., Rowen, L. & Kornberg, A. (1978) *J. Biol. Chem.*, in press.
9. Scheckman, R., Weiner, A. & Kornberg, A. (1974) *Science* **186**, 987–993.
10. Wickner, S. & Hurwitz, J. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 4120–4124.
11. Ray, D. & Duebner, J. (1975) in *DNA Synthesis and Its Regulation*, eds. Goulian, M., Hanawalt, P. & Fox, C. F. (W. A. Benjamin, Menlo Park, CA), pp. 370–385.
12. Ray, D., & Duebner, J. (1977) *J. Mol. Biol.* **113**, 651–661.
13. Godson, G. N. (1977) *J. Mol. Biol.* **117**, 353–368.
14. Baas, P. D. & Jansz, H. S. (1972) *J. Mol. Biol.* **63**, 569–576.
15. Fiddes, J. C. (1976) *J. Mol. Biol.* **107**, 1–24.
16. Godson, G. N. & Boyer, H. (1974) *Virology* **62**, 270–275.
17. Greenfield, L., Simpson, L. & Kaplan, D. (1975) *Biochim. Biophys. Acta* **407**, 365–375.
18. Jeppesen, P. G. N. (1974) *Anal. Biochem.* **58**, 195–207.
19. Peacock, A. C. & Dingman, C. W. (1968) *Biochemistry* **7**, 668–674.
20. Galibert, F., Sedat, J. & Ziff, E. B. (1974) *J. Mol. Biol.* **87**, 377–407.
21. Sanger, F. & Coulson, A. R. (1975) *J. Mol. Biol.* **94**, 441–448.
22. Barrell, B. G. (1978) in *Biochemistry of Nucleic Acids*, ed. Clark, B. F. C., International Review of Biochemistry, Series II (MTP Press, Lancaster, England), in press.
23. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA.* **74**, 5463–5467.
24. Godson, G. N. (1974) *Virology* **63**, 320–335.
25. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., III, Slocombe, P. M. & Smith, M. (1977) *Nature* **265**, 687–695.
26. Baas, P. D., Jansz, H. S. & Sinsheimer, R. L. (1976) *J. Mol. Biol.* **102**, 633–656.
27. Godson, G. N. (1974) *J. Mol. Biol.* **90**, 127–141.
28. Langeveld, S. A., van Mansfeld, A. D. M., Baas, P. D., Hansz, H. S., van Arkel, G. A. & Weisbeek, P. J. (1978) *Nature*, in press.
29. Eisenberg, S., Harbers, B., Hours, C. & Denhardt, D. T. (1975) *J. Mol. Biol.* **99**, 107–123.
30. Denniston-Thompson, K., Moore, D. D., Kruger, K. E., Furth, M. E. & Blattner, F. R. (1977) *Science*, **198**, 1051–1056.
31. Barrell, B. G., Air, G. M. & Hutchison, C. A., III (1976) *Nature* **265**, 34–41.
32. Eisenberg, S., Griffiths, J. & Kornberg, A. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3198–3202.
33. Tomizawa, J., Ohmori, H. & Bird, R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 1865–1869.
34. Ravetch, J. V., Horiuchi, K. & Zinder, N. D. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 4219–4222.