JOURNAL OF VIROLOGY, Mar. 1988, p. 882–886
0022-538X/88/030882-05$02.00/0

Vol. 62, No. 3

Nucleotide Sequence of the Tail Tube Structural Gene of Bacteriophage T4

FUMIO ARISAKA,[1]* LANCE ISHIMOTO,[2] GEORGE KASSAVETIS,[3] TAKASHI KUMAZAKI,[1] AND SHIN-ICHI ISHII[1]

Department of Biochemistry, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan 060[1]; Department of Microbiology and Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California 90024-1489[2]; and Department of Biology, University of California, San Diego, La Jolla, California 92093[3]

Received 3 August 1987/Accepted 28 October 1987

The nucleotide sequence of gene 19 of bacteriophage T4, the structural gene of the tail tube protein, was determined by both the dideoxy and the Maxam-Gilbert methods. The predicted M_r of tube protein gene product 19 is 18,842. The N-terminal amino acid of the tube protein was determined by Edman degradation, and the C-terminal sequence was confirmed by isolation of the C-terminal tryptic peptide. In the noncoding region between genes 18 and 19, there are two late-T4-promoter consensus sequences, 51 bases apart. The implication of the two late promoter sequences was examined by an S1 nuclease protection experiment. Both serve as weak promoters, but the bulk of the transcripts arise from further upstream of the two promoters.

The contractile tail of bacteriophage T4 consists of three parts: the contractile sheath, the tube, and the baseplate. The sheath protein is coded for by T4 gene 18, the tube protein is coded for by gene 19, and the baseplate is made up of 18 different proteins (2). Gene product (gp) 19 is a dimorphic protein that is made in an assembly-naive monomeric state in the infected cell. During T4 tail assembly, the presence of a complete T4 baseplate triggers a conformational change in gp19 that makes it competent to begin to polymerize into a tail tube; subsequent interactions with the "switched" gp19 continues the process (5). As the T4 tail tube polymerizes on the baseplate, it is surrounded by copolymerizing sheath monomers, eventually resulting in a complete T4 tail after the addition of gp3 and gp15 (5, 9). Both the tube and the sheath consist of 144 protomers arranged in 24 annuli. Upon infection, the sheath contracts and the tube penetrates the outer membrane of the host *Escherichia coli*. The T4 genome is then transferred from the head through the tube into the periplasmic space of the cell. During the infection process, as well as in the morphogenesis of the phage, tube protein gp19 must interact with several other components of the T4 tail. Some gp19 is in contact with the baseplate, and all of it is in contact with the sheath protein before infection. These interactions are rearranged when the sheath contracts. The tube protein also interacts with the length-determining protein in the center of the tube (4). As a step toward understanding the nature of the interactions in these processes, we have determined the predicted amino acid sequence of gp19 by determining the sequence of the gene. The deduced amino acid sequence of gp19 was then confirmed both by the N-terminal analysis of gp19 and isolation of the C-terminal tryptic peptide. In addition, the nucleotide sequence of gene 19 revealed that there are two T4 phage late-promoter consensus sequences upstream of the initiation codon of gene 19. This was surprising in view of the strong polarity effects of amber mutations in gene 18 on gene 19 expression. S1 nuclease protection experiments were performed to determine the role of the promoter sequences. The results are discussed with regard to these polarity effects.

* Corresponding author.

MATERIALS AND METHODS

Bacteria and bacteriophage. The *E. coli* strains used were from our collection and included RR1 (for plasmid construction and maintenance) and JM103 (as a host for the recombinant M13 phage used in the dideoxy sequencing). Messing M13 vectors mp8 and mp9 were used to generate subclones for DNA sequencing (14). *E. coli* Be was used as a *sup*^{0} indicator bacterial strain and for marker rescue spot tests. T4 gene 19 amber mutant NG524 was used to identify gene-19-containing clones and subclones in marker rescue tests.

Recombinant plasmid and phage construction and DNA manipulations. Construction of the plasmid clone containing the gene 19 region is described elsewhere (J. D. Conway, Ph.D. thesis, University of California, Los Angeles, Los Angeles, 1986). Large scale plasmid purification was performed as described by Godson and Vapnek (6). Restriction endonuclease digestion, DNA electroelution, ligation of DNA, and DNA transformations and transfections were done as described by Maniatis et al. (11) and by Messing and Viera (14).

Marker rescue test. Recombinant M13 clones were characterized by a simple marker rescue spot test (when possible) essentially as described by Mattson et al. (12).

DNA sequencing. The nucleotide sequence of gene 19 clones in plasmid p664 (12) was determined by both the method of Maxam and Gilbert (13) and the dideoxy method of Sanger et al. (16) as modified by Biggin et al. (3). The sequencing scheme used is diagrammed in Fig. 1. Sequencing by the dideoxy method covered all of gene 19 in both directions and confirmed the nucleotide sequence determined by the Maxam-Gilbert method.

Structural analysis of gp19. GP19 was isolated from tube baseplates by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and subjected to Edman degradation to determine the N-terminal amino acid sequence (7). The resulting amino acid derivative was converted to the phenylthiohydantoin amino acid, which was analyzed isocrati-

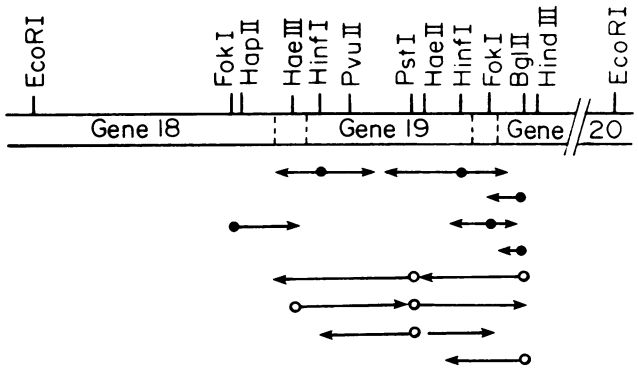

FIG. 1. Organization of the gene 18-19-20 region and the sequencing strategy for gene 19. With a variety of restriction endonuclease sites, DNA fragments were either subjected to base-specific chemical degradation or cloned into mp8 or mp9 and sequenced by the dideoxy method. Symbols: →, R-strand sequence determination; ←, L-strand sequence data; ●, sequence information from the Maxam-Gilbert method: ○, dideoxy sequencing with M13 clones.

cally by high performance liquid chromatography on an ODS column with a solvent system of 34% CH_3CN and 25 mM sodium acetate (pH 4.5) at 45°C. The C-terminal peptide was isolated from a tryptic digest of gp19 as described by Kumazaki et al. (10). Amino acid compositions of the C-terminal peptides, as well as of the whole protein, were determined after hydrolysis with 4 M methanesulfonic acid at 115°C for 24 h on an Hitachi 835 amino acid analyzer by the method of Simpson et al. (17).

S1 nuclease protection analysis. S1 protection experiments were carried out by the method of Kassavetis and Geiduschek (8). A *Pvu*II-*Pvu*II fragment containing these promoter sequences was prepared from plasmid p664/pTL12. The 5′ ends of this fragment were ^{32}P labeled, and the fragment was then cut with *Fnu*DII (see Fig. 5), hybridized with mRNA isolated from *E. coli* 18 min after T4 infection, treated with S1 nuclease, and run into a denaturing polyacrylamide gel and autoradiographed.

RESULTS

Sequence determination of gene 19. The sequencing scheme is shown in Figure 1. The whole sequence was determined by both the dideoxy and the Maxam-Gilbert sequencing methods except for a small region (approximately 40 base pairs [bp]), where the sequence was determined only by the dideoxy method. The dideoxy method was applied throughout the sequence in both directions and confirmed the sequence determined by the Maxam-Gilbert method. The nucleotide sequence thus determined is shown in Fig. 2, where the deduced amino acid sequence is also listed.

Description of the nucleotide sequence. The DNA sequence which has been determined covers the whole of gene 19, the 3′ end of gene 18, and the 5′ end of gene 20 (Fig. 2). Between genes 18 and 19 there is a 119-bp noncoding region. Two consensus T4 late-promoter sequences, TATAAATA, are present in the intergenic region and are separated by 51 nucleotides. A putative ribosome-binding site is also present at position 157. At the 5′ end, loop structures with respective stabilization energies of −19.4 and −15.4 kcal (−81,200 and −64,400 J, respectively) are predicted (15). A stem-and-loop structure with a calculated free energy of −21.2 kcals (88,700 J) is also predicted in the 3′-end region of gene 19. This potential transcription terminator, as well as the previously mentioned stem-and-loop structures, are shown in Fig. 3.

Sequence analyses of gp19. The N-terminal amino acid sequence analysis of gp19 (see the legend to Table 1) confirmed the reading frame predicted by the nucleotide sequence and demonstrated that Met is the N-terminal amino acid. The amino acid composition analysis of the C-terminal tryptic peptide isolated by specific adsorption of the other tryptic peptides from gp19 on anhydrotrypsin-Sepharose also confirmed the C-terminal amino acid sequence deduced from the DNA sequence data (10). Gp19

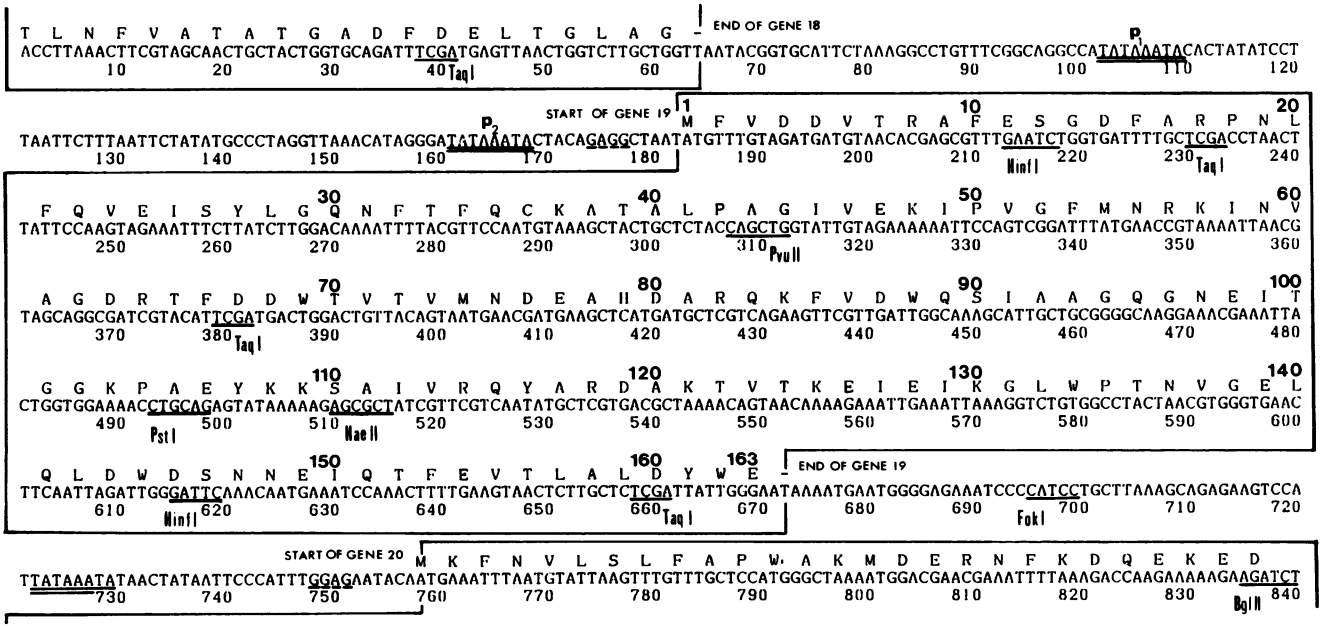

FIG. 2. DNA sequence of the gene 19 region and the predicted amino acid sequence of gp19. The consensus T4 late promoters and a putative Shine-Dalgarno sequence are indicated by underlining: - - - - -, S/D sequence; =, T4 late-promoter consensus sequence.

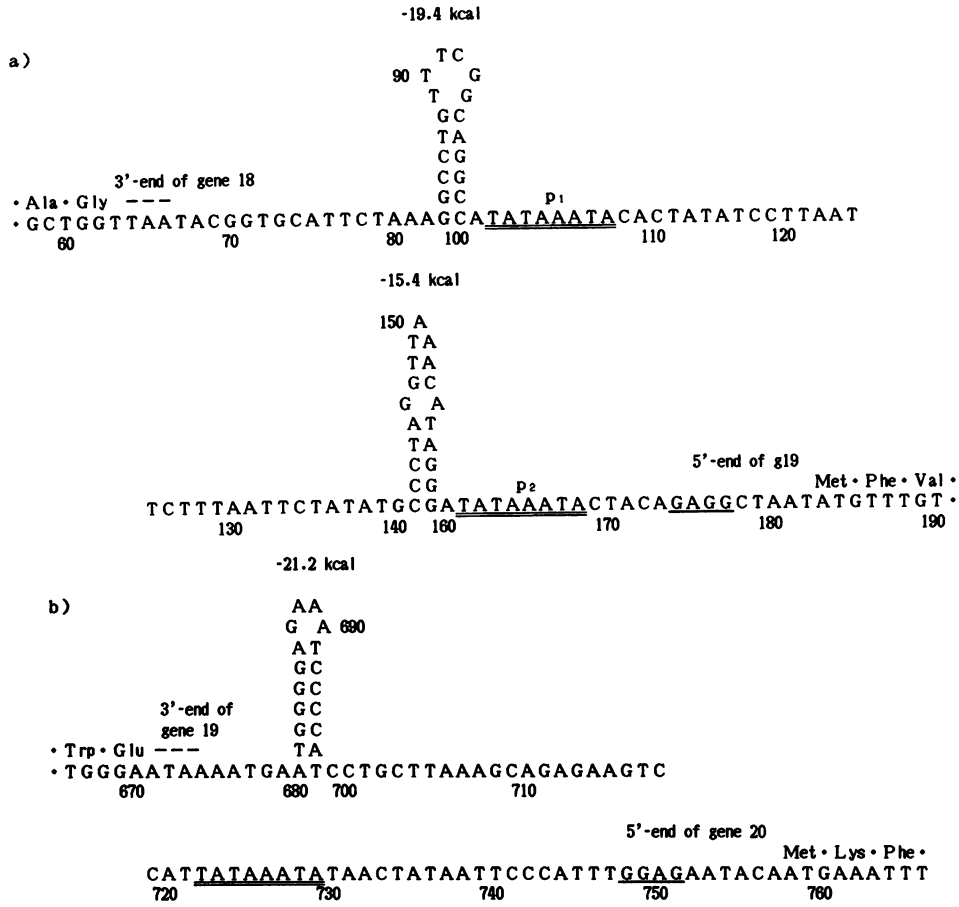

FIG. 3. Stem-and-loop structures in the gene 18-19 intergenic region (a) and distal to gene 19 (b). Three theoretical stem-and-loop structures and their respective calculated free energies are shown. The function of the two secondary structures that are adjacent to the two consensus T4 late-promoter regions (underlined) is unknown. The stem-and-loop structure at the 3' end of gene 19 functions as a transcription terminator.

was thus concluded to consist of 163 amino acid residues (Fig. 2) (M_r, 18,842). The amino acid composition of the whole protein was in excellent agreement with that expected from the nucleotide sequence except for glycine, whose disparate values were due to contamination from the sodium dodecyl sulfate-polyacrylamide gel (Table 1).

S1 nuclease protection analysis of the two late-promoter sequences between genes 18 and 19. To test whether the two consensus T4 late promoters found in the gene 18-19 intergenic region are used in vivo, S1 nuclease protection analysis was performed. The results (Fig. 4) demonstrated clearly that both of the intergenic promoters function in vivo, although the vast majority of gene-19-containing transcript appears to originate upstream from the two promoters. The high background seen in lane 1 (minus S1) is due to the fact that S1 protection experiments were carried out with an excess of ^{32}P-labeled DNA. The major bands seen are the 491- and 60-bp bands that correspond to the two end-labeled *Pvu*II-*Pvu*II fragments (Fig. 5); the other bands are ^{32}P-labeled artifacts resulting from the DNA purification and end-labeling procedures. The end-labeled DNA, treated with S1 nuclease in the absence of T4 late mRNA, is shown in Fig. 4, lane 2. The only two bands seen are the 491- and 60-bp *Pvu*II-*Fnu*DII fragments, which have been reannealed and become insensitive to S1 activity. Lanes 4 through 7 show S1 reactions to which increasing amounts of T4 mRNA were added. Protected regions of 196 and 138 nucleotides,

TABLE 1. Amino acid composition of gp19[a]

Amino acid	No. of residues found (predicted)
Asx	21.0 (13 [Asp], 9 [Asn])
Thr	10.8 (12)
Ser	8.0 (5)
Glx	21.2 (12 [Glu], 9 [Gln])
Pro	5.9 (5)
Gly	16.6 (11)
Ala	15.0[b] (15)
Cys	0.6 (1)
Val	10.9 (13)
Met	1.9 (3)
Ile	8.1 (10)
Leu	8.0 (8)
Tyr	4.2 (4)
Phe	8.9 (10)
His	1.5 (1)
Lys	10.2 (10)
Trp	3.4 (5)
Arg	7.6 (7)

[a] See Materials and Methods. An N-terminal analysis of gp19 showed the following phenylthiohydantoin amino acid sequence (and content [in nanomoles]): Met (3.9)-Phe (3.4)-Val (5.1)-Asp (8.7)-Asp (9.6)-Val (6.4).

[b] Ala was taken as the standard.

corresponding to transcripts initiating at both intergenic promoters, are visible in lanes 6 and 7. The 491-nucleotide protected region can be seen to increase in intensity in lanes 4 through 7. This increase is due to the fact that the majority of transcripts capable of protecting the labeled DNA strand originate from a region upstream from gene 19, presumably at the gene 18 promoter. The S1 protection experiments, which involved the entire 749-bp *Pvu*II fragment (Fig. 5) that was ^{32}P end labeled and strand separated, verified this interpretation (data not shown). In these experiments, no labeled DNA survived S1 treatment in the absence of T4 mRNA and results similar to those seen in Fig. 4, lanes 4 through 7, were obtained with the exception that all of the bands seen on the autoradiogram were due to T4 mRNA-mediated S1 protection.

DISCUSSION

The nucleotide sequence of gene 19 and its respective product, which have been described, contain no discernible internal repeats, and gp19 shares no amino acid homology

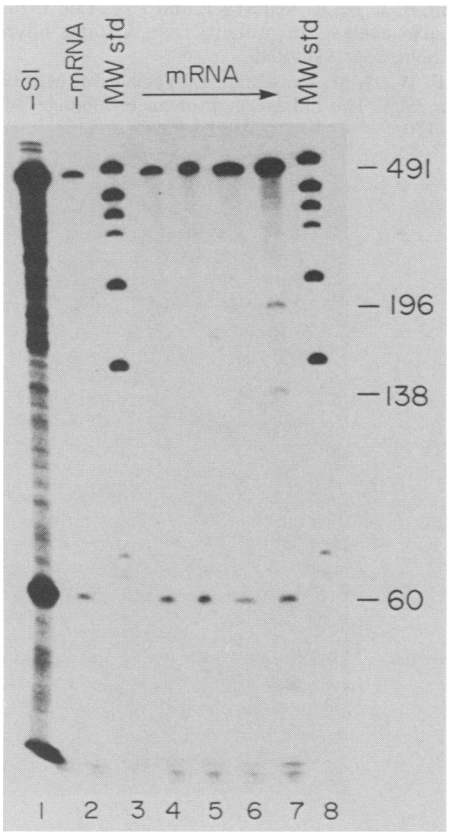

FIG. 4. Autoradiogram showing the T4 late-mRNA-mediated S1 nuclease protection pattern of the 491-nucleotide ^{32}P-end-labeled DNA strand described in Fig. 5. ^{32}P-end-labeled DNA (1 pmol) containing the gene 18-19 intergenic region was heat denatured, annealed with T4 late mRNA, digested with S1 nuclease, and run onto a 4% polyacrylamide gel in the presence of 7 M urea. Shown are controls to which no S1 nuclease was added after annealing (lane 1) and to which no T4 late mRNA was added before S1 digestion (lane 2); molecular weight standards (MW std) (lanes 3 and 8); and 0.2 (lane 4), 0.6 (lane 5), 1.5 (lane 6), and 5.0 (lane 7) μg of total T4 RNA from T4-infected cells annealed to the ^{32}P-end-labeled DNA fragments. The resulting increase in the 491-, 196-, and 138-nucleotide end-labeled DNA protection products can be seen.

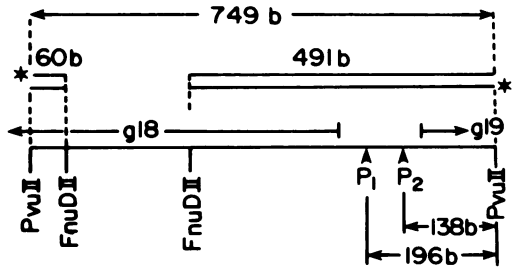

FIG. 5. Strategy used to create ^{32}P-end-labeled DNA probes for S1 nuclease protection studies. A 749-bp *Pvu*II fragment was ^{32}P end labeled (*) and then cut with *Fnu*DII, resulting in two double-stranded ^{32}P-end-labeled fragments that were 60 and 491 bp (b) in length. The relation of these fragments to the gene 18-19 intergenic region is indicated. The two consensus T4 promoters in the intergenic region are labeled P_1 and P_2. Transcripts originating at P_1 or P_2 would partially protect the 491-nucleotide ^{32}P-end-labeled DNA strand from S1 nuclease digestion, resulting in 196- and 138-nucleotide ^{32}P-end-labeled DNA strands. Transcripts originating upstream from the *Fnu*DII site would protect the entire 491-nucleotide ^{32}P-end-labeled DNA strand.

with the tail sheath protein, gp18 (F. Arisaka, T. Nakako, H. Takahashi, and S.-I. Ishii, J. Virol., in press) of phage T4, or the major tail protein of bacteriophage lambda, gpV. GP19, however, does have a significant amino acid homology with a part of gp54 of phage T4 (Ishimoto et al., submitted for publication). GP54 is a baseplate protein with an M_r of 36,000 that appears to function as a tail tube polymerization initiator (9). In light of the proposed function of gp54, the homology between it and gp19 is especially interesting. It is very likely that the homologous region in gp54 is involved in gp19-gp54 or gp19-baseplate interaction and that the homologous region in gp19 is the interface between gp19 protomers in the tail tube.

N-terminal Met is retained in the mature gene product. This is in agreement with the report of Ben-Bassat and Bauer (1) that N-terminal Met is generally retained when residue 2 is phenylalanine. The removal of N-terminal Met in the case of the tail sheath protein, gp18, also agrees with their rule, in which residue 2 was threonine (Arisaka et al., in press).

S1 protection experiments have shown that the two gene 19-proximal late-promoter sequences served as weak promoters, but the bulk of the transcripts coding for gene 19 originated further upstream of the promoters, possibly from the promoter for gene 18. The result explains why there is a strong polarity effect between genes 18 and 19 (20 to 40% [18]), yet the suppression of gp19 production, when gene 18 has an amber mutation, is not complete. The result also implied that the two stem-loop structures present in the noncoding region between genes 18 and 19 (Fig. 3) were at least not strong terminators or attenuators for transcription. However, the significance of the two promoters for phage growth or any other advantage to the phage is not certain.

ACKNOWLEDGMENTS

The authors thank Frederick A. Eiserling and E. Peter Geiduschek for valuable discussions and comments on the manuscript.

This work was supported in part by grants-in-aid for scientific research from the Ministry of Education, Science, and Culture of Japan and by Public Health Service grants AI 14092 to F.A.E. and GM 18386 to E.P.G. from the National Institutes of Health.

LITERATURE CITED

1. **Ben-Bassat, A., and K. Bauer.** 1987. Amino-terminal processing of proteins. Nature (London) **326**:315.
2. **Berget, P. B., and J. King.** 1983. T4 tail morphogenesis, p. 246–258. *In* C. K. Matthews, E. M. Kutter, G. Mosig, and P. B. Berget (ed.), Bacteriophage T4. American Society for Microbiology, Washington, D.C.
3. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80**:3963–3965.
4. **Duda, R. L., J. S. Wall, J. F. Hainfeld, R. M. Sweet, and F. A. Eiserling.** 1985. Mass distribution of a probable tail-length-determining protein in bacteriophage T4. Proc. Natl. Acad. Sci. USA **82**:5550–5554.
5. **Eiserling, F. A.** 1983. Structure of the T4 virion, p. 11–24. *In* C. K. Matthews, E. M. Kutter, G. Mosig, and P. B. Berget (ed.), Bacteriophage T4. American Society for Microbiology, Washington, D.C.
6. **Godson, G. N., and D. Vapnek.** 1973. A simple method of preparing large amounts of ϕx174 RFI super-coiled DNA. Biochim. Biophys. Acta **299**:516–521.
7. **Hunkapiller, M. W., E. Lujan, F. Ostrander, and L. E. Hood.** 1983. Isolation of microgram quantities of proteins from polyacrylamide gels for amino acid sequence analysis. Methods Enzymol. **91**:227–236.
8. **Kassavetis, G. A., and E. P. Geiduschek.** 1982. Bacteriophage T4 late promoters: mapping 5′ ends of T4 gene 23 mRNAs. EMBO J. **1**:107–114.
9. **King, J.** 1971. Bacteriophage T4 tail assembly: four steps in core formation. J. Mol. Biol. **58**:693–709.
10. **Kumazaki, T., T. Nakako, F. Arisaka, and S. Ishii.** 1986. A novel method for selective isolation of C-terminal peptides from tryptic digests of proteins by immobilized anhydrotrypsin: application to structural analyses of the tail sheath and tube proteins from bacteriophage T4. Proteins: Struct., Function, Genet. **1**:100–107.
11. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
12. **Mattson, T., G. V. Houwe, A. Bolle, G. Selzer, and R. Epstein.** 1977. Genetic identification of cloned fragments of bacteriophage DNA and complementation by some clones containing early T4 genes. Mol. Gen. Genet. **154**:319–326.
13. **Maxam, A., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavage. Methods Enzymol. **65**:499–560.
14. **Messing, J., and J. Viera.** 1982. A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene **19**:269–276.
15. **Ooi, T., and M. Takanami.** 1981. A computer method for construction of secondary structure from polynucleotide sequence: possible structure of the bacterial replication origin. Biochim. Biophys. Acta **655**:221–229.
16. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain termination inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
17. **Simpson, R. J., M. R. Neuberger, and T.-Y. Liu.** 1976. Complete amino acid analysis of proteins from a single hdyrolysate. J. Biol. Chem. **251**:1936–1940.
18. **Stahl, F. W., J. M. Craseman, C. Yegian, M. M. Stahl, and A. Nakata.** 1964. The linkage map of bacteriophage T4. Genetics **64**:157–170.