Phage P22 Tail Protein: Gene and Amino Acid Sequence[†]

Robert T. Sauer,* William Krovatin, Anthony R. Poteete, and Peter B. Berget

ABSTRACT: The tail structure of the *Salmonella* phage P22 mediates both adsorption of the phage to its host and enzymatic hydrolysis of the bacterial O-antigen. The tail is an oligomeric structure, which is assembled from a single polypeptide species. We report here the amino- and carboxyl-terminal sequences of the P22 tail protein and the nucleotide sequence of its gene (gene 9). These data specify the complete amino acid sequence of the tail protein. The tail protein is a slightly acidic protein containing 666 amino acids. Comparison of the gene and protein sequences indicates that mature tail protein arises by cleavage of the initiator *N*-formylmethionine from the nascent chain.

Virions of the *Salmonella* phage P22 consist of an icosahedral protein head, containing the packaged phage DNA, and a short tail, composed of a single polypeptide species. This tail polypeptide is encoded by gene 9 of P22. During morphogenesis, tail protein is the last structural protein to assemble onto the virus. It is present during the entire assembly period but only interacts with otherwise fully completed heads into which the phage chromosome has been packaged (Anderson, 1960; Israel et al., 1967; Botstein et al., 1973; King et al., 1973). Prior to assembly, or in the absence of other phage proteins, tail protein is trimeric (Goldenberg, 1981). The tail structure of the virion contains six of these trimers. Tail protein mediates binding of the virus to susceptible cells (Israel et al., 1967; Botstein et al., 1973) and also catalyzed specific cleavage of rhamnosyl-1,3-galactose linkages in the bacterial O-antigen (Iwashita & Kanegasaki, 1973, 1976; Eriksson & Lindberg, 1977).

P22 tail protein, thus, participates in four distinct interactions: trimer formation, assembly onto phage heads, binding to susceptible cells, and cleavage of specific sugar linkages. The assembly, absorption, and cleavage reactions proceed efficiently in purified systems and thus are accessible to biochemical study (Berget & Poteete, 1980). Moreover, although gene 9 is essential for lytic growth, phage bearing mutations in gene 9 can be propagated in medium supplemented with tail protein. This property of the system makes gene 9 conditionally essential and thus makes the individual activities of the tail protein extraordinarily accessible to genetic study. Because tail protein functions both as an enzyme and as a structural protein, and since these activities can be studied both biochemically and genetically, tail protein provides an attractive model system in which to study macromolecular structure and function. To aid in such studies, we have cloned and sequenced the structural gene for the P22 tail protein. We report here the DNA sequence of gene 9 and the inferred amino acid sequence of the tail protein.

Materials and Methods

Tail Protein Purification. P22 tail protein was purified from *Escherichia coli* strain MM294 bearing the tail-producing plasmid pPB13 (P. Berget, A. Poteete, and R. T. Sauer, unpublished results). Cells were grown at 37 °C in AZ broth (Ueda et al., 1978) to stationary phase in a 200-L fermentor. The cell paste obtained after continuous flow centrifugation was resuspended in an equal weight of B_{25} buffer (Berget & Poteete, 1980) and was frozen by pouring into liquid nitrogen. Frozen cells were stored at −80 °C; 900 g of frozen cell suspension was thawed and allowed to warm to room temperature after addition of 900 mL of B_{25} buffer. The cell suspension was mixed and brought to pH 8, 15 mM EDTA[1] and 1 mM DTT by addition of 2 M Tris base, 0.5 M EDTA (pH 8) and 0.5 M DTT, respectively. Crystalline egg white lysozyme was added to 200 μg/mL, and the suspension was immediately transferred to Sorvall GSA bottles. These were then incubated at 37 °C for 30 min, at which time lysis was apparent by the increase in viscosity. The cell lysate was centrifuged at 12 000 rpm at 4 °C for 1 h. The clear amber supernatant was decanted, warmed to room temperature, and adjusted to pH 4 by addition of glacial acetic acid. At this point a heavy precipitate formed. The acidified extract was heated to 65 °C in a shaking water bath and held at that temperature for 30 min. Following chilling to 0 °C in an ice bath, the suspension was clarified by centrifugation at 12 000 rpm, at 4 °C, in the GSA rotor of a Sorvall centrifuge. The resulting supernatant fraction was adjusted to pH 7.6 by addition of 2 M Tris base, and tail protein was precipitated by bringing the solution to 40% saturation with ammonium sulfate. Precipitated material was resuspended in 100 mL of 50 mM sodium acetate (pH 4) and was dialyzed overnight, at room temperature, against 40 volumes of the same buffer. The cloudy dialyzed sample was clarified by centrifugation at 20 000 rpm in the Sorvall SS-34 rotor at 4 °C. The resulting supernatant fraction was approximately 90–95% pure tail protein as determined by NaDodSO₄–polyacrylamide gel electrophoresis. The minor contaminating proteins were removed by chromatography of the sample on a 275-mL column of SP-Sephadex (A-50) equilibrated with 50 mM sodium acetate (pH 4). The tail protein eluted from this column roughly in the middle of a 4-L gradient from 0 to 0.5 M NaCl in the equilibration buffer. At this point, the tail protein was homogeneous by NaDodSO₄–polyacrylamide gel electrophoresis. Purified tail protein was stored at 4 °C after dialysis into 50 mM sodium acetate (pH 4) buffer. The final yield of purified tail protein was approximately 0.4 mg/g of cells.

Some initial studies of tail protein were performed on protein purified as described (Berget & Poteete, 1980; Goldenberg

[1] Abbreviations: EDTA, ethylenediaminetetraacetic acid; DTT, dithiothreitol; Tris, tris(hydroxymethyl)aminomethane; NaDodSO₄, sodium dodecyl sulfate.

& King, 1981) and provided by D. Goldenberg.

Nucleic Acid Chemistry. Restriction fragments bearing portions of P22 gene 9 were purified either from plasmid pPB10 (P. Berget, A. Poteete, and R. Sauer, unpublished results) or from phage P22 Ap31 pfr1 DNA (Winston, 1980) by conventional procedures (Maniatis et al., 1975; Maxam & Gilbert, 1980). Restriction fragments with 5′-protruding termini were 3′ end labeled by using the large fragment of *E. coli* DNA polymerase I (Klenow & Henningsen, 1970; Setlow & Kornberg, 1972) and [α-^{32}P]deoxyribonucleotide 5′-triphosphates. Singly end-labeled restriction fragments were obtained either by secondary restriction cleavage or by strand separation (Maxam & Gilbert, 1980). DNA sequencing was performed by the chemical method of Maxam & Gilbert (1977, 1980). Initial restriction mapping of the P22 gene 9 region with *Hae*III, *Hinf*I, *Alu*, *Hpa*II, and *Hha*I was performed by partial restriction endonuclease digestion (Smith & Birnstiel, 1976) of fragments singly end labeled at the *Eco*RI site in the P22 *ant* gene or the *Bam*H1 site in gene 9 (Figure 1). In some cases, the restriction map deduced from this analysis was confirmed by double-restriction digests. Our initial restriction map of the gene 9 region was used to direct sequencing strategy, but in the final primary gene sequence, partial sequences from specific restriction fragments were aligned by direct sequence overlap of at least 15 bases.

Protein Chemistry. Samples of purified tail protein were hydrolyzed for 24, 48, 72, or 96 h in 5.7 N HCl plus 1% phenol at 110 °C in vacuo. Amino acid analyses were performed on a Durham D500 analyzer. For calculation of the extinction coefficient of the tail protein, samples were dissolved in 50 mM ammonium acetate (pH 4), their UV absorbance was scanned, then aliquots were dried under vacuum and hydrolyzed, and the protein concentration was determined by amino acid analysis. The experimentally determined molar extinction coefficient at 278 nm was 7.3 × 10^4 mol^{-1} cm^{-1} L.

Automated Edman degradations were performed on a Beckman 890C sequencer using the 0.1 M Quadrol program described by Brauer et al. (1975). 2-Anilinothiazolinone derivatives were converted to phenylthiohydantoin amino acid derivatives in 1 N HCl (80 °C, 10 min) (Ilse & Edman, 1960). PTH-amino acid derivatives were identified by gas–liquid, high-performance liquid, and thin-layer chromatography using procedures described previously (Sauer & Anderegg, 1978; Sauer et al., 1981).

Carboxypeptidase Y digestions were carried out at 37 °C (E/S = 1/500 mol/mol) in a buffer containing 1% triethylamine–acetate (pH 6.5) and 1% sodium dodecyl sulfate. Prior to digestion, tail protein was suspended in 1% triethylamine-acetate (pH 6.5) and 4% NaDodSO$_4$ and was boiled for 3 min. Amino acids released by carboxypeptidase digestion were analyzed directly by high-performance liquid chromatography after derivitization with *o*-phthalaldehyde and ethanethiol (Hill et al., 1979). Fluorescent amino acid derivatives were separated on a Waters μBondapak C$_{18}$ column. In this system, all amino acids except proline are separated and detected. Limit digestions were also analyzed by conventional amino acid analysis to check for the presence of proline and to confirm the results obtained by HPLC. In these cases, Asn, Gln, and Ser were not separated.

Results

DNA Sequence of the Gene 9 Region. To precisely locate gene 9 coding sequences, we sequenced approximately 5% of the P22 chromosome (map coordinates 0.38–0.43). Previous genetic studies had indicated that at least part of gene 9 was contained in this region of DNA (Weinstock, 1977; Chisholm et al., 1980; P. B. Berget, unpublished results). The positions
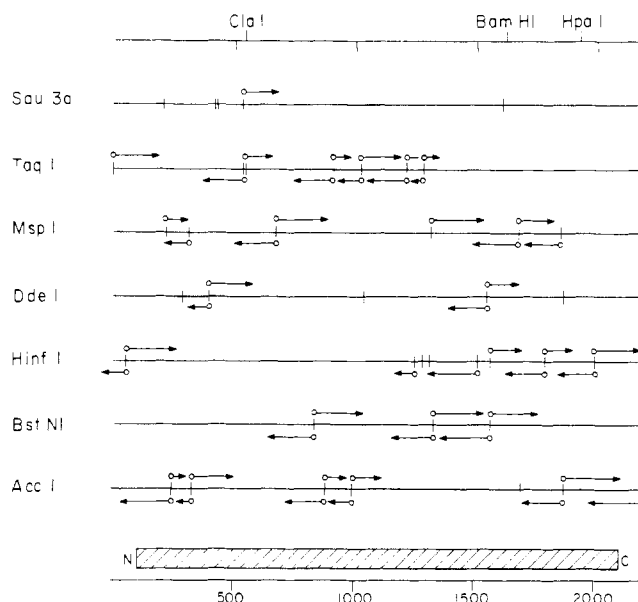

FIGURE 1: DNA sequencing strategy. Restriction sites for *Sau*3a (GATC), *Taq*I (TCGA), *Msp*I (CCGG), *Dde*I (CTNAG), *Hinf*I (GANTC), *Bst*NI (CC^A_TGG), *Acc*I (GT$^{AG}_{CT}$AC), *Cla*I (ATCGAT), *Bam*HI (GGATCC), and *Hpa*I (GTTAAC) are indicated by vertical lines. Circles indicate positions of 3′ end labeling. Arrows indicate extent of reliable DNA sequence obtained. The bar represents gene 9 coding sequences.

of specific restriction fragments used for the sequencing and the extent of reliable sequence information obtained from each set of sequencing reactions are shown schematically in Figure 1. Plasmid pPB10 was the source of gene 9 DNA for most of the sequence experiments reported here. pPB10 carries approximately 4.0 kilobases of P22 DNA (extending from the *Eco*RI site in the P22 *ant* gene to the *Pvu*II site downstream from gene 9) cloned between the *Eco*RI and *Pvu*II sites of pBR 322 (P. B. Berget, A. Poteete, and R. T. Sauer, unpublished results). In one case, phage P22 Ap31 pfr1 was the source of DNA. Direct overlapping of DNA sequences obtained from individual restriction fragments resulted in a unique continuous sequence. This sequence is shown in Figure 2.

The restriction endonuclease sites predicted from the DNA sequence (Figure 2) were consistent with mapping experiments with one exception. The $\begin{array}{c}5'\text{-ATCGATCGA-}3'\\3'\text{-TAGCTAGCT-}5'\end{array}$ sequence at bases 547–555 should encode two tandem cleavage sites for the *Taq*I endonuclease (recognition sequence TCGA), as well as single sites for the *Sau*3a (GATC), *Pvu*I (CGATCG), and *Cla*I (ATCGAT) endonucleases. The N6 positions of the adenines in the *Sau*3a site are potential sites of *dam*-mediated methylation (Marinus & Morris, 1973; Geier & Modrich, 1979). When gene 9 DNA was obtained from plasmid pPB10 grown in *E. coli* strain 294 (*dam*$^+$), cleavage at *Sau*3a and *Pvu*I sites was observed, but limited cleavage at the tandem *Taq* sites and no cleavage at the *Cla* site were obtained. In contrast, cleavage at all four sites could be obtained with plasmid DNA grown in *dam*$^-$ strains of *E. coli*. Methylation of the adenine in the *Taq* recognition sequence is known to inhibit *Taq* cleavage (Backman, 1980), and our results suggest that adenine methylation also prevents cleavage by the *Cla*I endonuclease. Efficient cleavage at the tandem *Taq* sites and the *Cla* site was obtained when phage P22 (grown on *Salmonella typhimurium* strain DB 7000) was used as the source of gene 9 DNA. We infer either that this host strain is *dam*$^-$ or that P22 growth prevents adenine modification either by encoding an inhibitor or by overwhelming the host modification

****

```
AAATTGTAGGTCGAA AAGCGAACAAAATAA CTTCCGAAAAAGTTG TTTTTATCACAAAAAA TTCACCGTAGCCATG CTGCGGCAATTCCTT GCATCTGGAGCAAAT TAA ATG

                    5                    10                   15                   20                   25                   30
THR ASP ILE THR ALA ASN VAL VAL VAL SER ASN PRO ARG PRO ILE PHE THR GLU SER ARG SER PHE LYS ALA VAL ALA ASN GLY LYS ILE
ACA GAC ATC ACT GCA AAC GTA GTT GTT TCT AAC CCT CGT CCA ATC TTC ACT GAA TCC CGT TCG TTT AAA GCT GTT GCT AAT GGG AAA ATT

                    35                   40                   45                   50                   55                   60
TYR ILE GLY GLN ILE ASP THR ASP PRO VAL ASN PRO ALA ASN GLN ILE PRO VAL TYR ILE GLU ASN GLU ASP GLY SER HIS VAL GLN ILE
TAC ATT GGT CAG ATT GAT ACC GAT CCG GTT AAT CCT GCC AAT CAG ATA CCC GTA TAC ATT GAA AAT GAG GAT GGC TCT CAC GTC CAG ATT

                    65                   70                   75                   80                   85                   90
THR GLN PRO LEU ILE ILE ASN ALA ALA GLY LYS ILE VAL TYR ASN GLY GLN LEU VAL LYS ILE VAL THR VAL GLN GLY HIS SER MET ALA
ACT CAG CCG CTA ATT ATC AAC GCA GCC GGT AAA ATC GTA TAC AAC GGC CAA CTG GTG AAA ATT GTC ACC GTT CAG GGT CAT AGC ATG GCT

                    95                   100                  105                  110                  115                  120
ILE TYR ASP ALA ASN GLY SER GLN VAL ASP TYR ILE ALA ASN VAL LEU LYS TYR ASP PRO ASP GLN TYR SER ILE GLU ALA ASP LYS LYS
ATC TAT GAT GCC AAT GGT TCT CAG GTT GAC TAT ATT GCT AAC GTA TTG AAG TAC GAT CCA GAT CAA TAT TCA ATA GAA GCT GAT AAA AAA

                    125                  130                  135                  140                  145                  150
PHE LYS TYR SER VAL LYS LEU SER ASP TYR PRO THR LEU GLN ASP ALA ALA SER ALA ALA VAL ASP GLY LEU LEU ILE ASP ARG ASP TYR
TTT AAG TAT TCA GTA AAA TTA TCA GAT TAT CCA ACA TTG CAG GAT GCA GCA TCT GCT GCG GTT GAT GGC CTT CTT ATC GAT CGA GAT TAT

                    155                  160                  165                  170                  175                  180
ASN PHE TYR GLY GLY GLU THR VAL ASP PHE GLY GLY LYS VAL LEU THR ILE GLU CYS LYS ALA LYS PHE ILE GLY ASP GLY ASN LEU ILE
AAT TTT TAT GGT GGA GAG ACA GTT GAT TTT GGC GGA AAG GTT CTG ACT ATA GAA TGT AAA GCT AAA TTT ATA GGA GAT GGA AAT CTT ATT

                    185                  190                  195                  200                  205                  210
PHE THR LYS LEU GLY LYS GLY SER ARG ILE ALA GLY VAL PHE MET GLU SER THR THR THR PRO TRP VAL ILE LYS PRO TRP THR ASP ASP
TTT ACG AAA TTA GGC AAA GGT TCC CGC ATT GCC GGG GTT TTT ATG GAA AGC ACT ACA ACA CCA TGG GTT ATC AAG CCT TGG ACG GAT GAC

                    215                  220                  225                  230                  235                  240
ASN GLN TRP LEU THR ASP ALA ALA ALA VAL VAL ALA THR LEU LYS GLN SER LYS THR ASP GLY TYR GLN PRO THR VAL SER ASP TYR VAL
AAT CAG TGG CTA ACG GAT GCC GCA GCG GTC GTT GCC ACT TTA AAA CAA TCT AAA ACT GAT GGG TAT CAG CCA ACC GTA AGC GAT TAC GTT

                    245                  250                  255                  260                  265                  270
LYS PHE PRO GLY ILE GLU THR LEU LEU PRO PRO ASN ALA LYS GLY GLN ASN ILE THR SER THR LEU GLU ILE ARG GLU CYS ILE GLY VAL
AAA TTC CCA GGA ATA GAA ACG TTA CTC CCA CCT AAT GCA AAA GGG CAA AAC ATA ACG TCT ACG TTA GAA ATT AGA GAA TGT ATA GGG GTC

                    275                  280                  285                  290                  295                  300
GLU VAL HIS ARG ALA SER GLY LEU MET ALA GLY PHE LEU PHE ARG GLY CYS HIS PHE CYS LYS MET VAL ASP ALA ASN ASN PRO SER GLY
GAA GTT CAT CGG GCT AGC GGT CTA ATG GCT GGT TTT TTG TTT AGA GGG TGT CAC TTC TGC AAG ATG GTA GAC GCC AAT AAT CCA AGC GGA

                    305                  310                  315                  320                  325                  330
GLY LYS ASP GLY ILE ILE THR PHE GLU ASN LEU SER GLY ASP TRP GLY LYS GLY ASN TYR VAL ILE GLY GLY ARG THR SER TYR GLY SER
GGT AAA GAT GGC ATT ATA ACC TTC GAA AAC CTT AGC GGC GAT TGG GGG AAG GGT AAC TAT GTC ATT GGC GGA CGA ACC AGC TAT GGG TCA

                    335                  340                  345                  350                  355                  360
VAL SER SER ALA GLN PHE LEU ARG ASN ASN GLY GLY PHE GLU ARG ASP GLY GLY VAL ILE GLY PHE THR SER TYR ARG ALA GLY GLU SER
GTA AGT AGC GCC CAG TTT TTA CGT AAT AAT GGT GGC TTT GAA CGT GAT GGT GGA GTT ATT GGG TTT ACT TCA TAT CGC GCT GGG GAG AGT

                    365                  370                  375                  380                  385                  390
GLY VAL LYS THR TRP GLN GLY THR VAL GLY SER THR THR SER ARG ASN TYR ASN LEU GLN PHE ARG ASP SER VAL VAL ILE TYR PRO VAL
GGC GTT AAA ACT TGG CAA GGT ACT GTG GGC TCG ACA ACC TCT CGC AAC TAT AAT CTG CAA TTC CGC GAC TCG GTC GTT ATT TAC CCC GTA

                    395                  400                  405                  410                  415                  420
TRP ASP GLY PHE ASP LEU GLY ALA ASP THR ASP MET ASN PRO GLU LEU ASP ARG PRO GLY ASP TYR PRO ILE THR GLN TYR PRO LEU HIS
TGG GAC GGA TTC GAT TTA GGT GCT GAC ACT GAC ATG AAT CCG GAG TTG GAC AGG CCA GGG GAC TAC CCT ATA ACC CAA TAC CCA CTG CAT

                    425                  430                  435                  440                  445                  450
GLN LEU PRO LEU ASN HIS LEU ILE ASP ASN LEU LEU VAL ARG GLY ALA LEU GLY VAL GLY PHE GLY MET ASP GLY LYS GLY MET TYR VAL
CAG TTA CCC CTA AAT CAC CTG ATT GAT AAT CTT CTG GTT CGC GGG GCG TTA GGT GTA GGT TTT GGT ATG GAT GGT AAG GGC ATG TAT GTG

                    455                  460                  465                  470                  475                  480
SER ASN ILE THR VAL GLU ASP CYS ALA GLY SER GLY ALA TYR LEU LEU THR HIS GLU SER VAL PHE THR ASN ILE ALA ILE ILE ASP THR
TCT AAT ATT ACC GTA GAA GAT TGC GCT GGG TCT GGC GCG TAC CTA CTC ACC CAC GAA TCA GTA TTT ACC AAT ATA GCC ATA ATT GAC ACC

                    485                  490                  495                  500                  505                  510
ASN THR LYS ASP PHE GLN ALA ASN GLN ILE TYR ILE SER GLY ALA CYS ARG VAL ASN GLY LEU ARG LEU ILE GLY ILE ARG SER THR ASP
AAT ACT AAG GAT TTC CAG GCG AAT CAG ATT TAT ATA TCT GGG GCT TGC CGT GTG AAC GGT TTA CGT TTA ATT GGG ATC CGC TCA ACC GAT

                    515                  520                  525                  530                  535                  540
GLY GLN GLY LEU THR ILE ASP ALA PRO ASN SER THR VAL SER GLY ILE THR GLY MET VAL ASP PRO SER ARG ILE ASN VAL ALA ASN LEU
GGG CAG GGT CTA ACC ATA GAC GCC CCT AAC TCT ACC GTA AGC GGT ATA ACC GGG ATG GTA GAC CCC TCT AGA ATT AAT GTT GCT AAT TTG

                    545                  550                  555                  560                  565                  570
ALA GLU GLU GLY LEU GLY ASN ILE ARG ALA ASN SER PHE GLY TYR ASP SER ALA ALA ILE LYS LEU ARG ILE HIS LYS LEU SER LYS THR
GCA GAA GAA GGG TTA GGT AAT ATC CGC GCT AAT AGT TTC GGC TAT GAT AGC GCA GCG ATT AAA CTG CGG ATT CAT AAG TTA TCA AAG ACA

                    575                  580                  585                  590                  595                  600
LEU ASP SER GLY ALA LEU TYR SER HIS ILE ASN GLY GLY ALA GLY SER GLY SER ALA TYR THR GLN LEU THR ALA ILE SER GLY SER THR
TTA GAT AGC GGA GCA TTG TAC TCC CAC ATT AAC GGG GGG GCC GGT TCT GGC TCA GCG TAT ACT CAA CTT ACT GCT ATT TCA GGT AGC ACA

                    605                  610                  615                  620                  625                  630
PRO ASP ALA VAL SER LEU LYS VAL ASN HIS LYS ASP CYS ARG GLY ALA GLU ILE PRO PHE VAL PRO ASP ILE ALA SER ASP ASP PHE ILE
CCT GAC GCT GTA TCA TTA AAA GTT AAC CAC AAA GAT TGC AGG GGG GCA GAG ATA CCA TTT GTT CCT GAC ATC GCG TCA GAT GAT TTT ATA

                    635                  640                  645                  650                  655                  660
LYS ASP SER SER CYS PHE LEU PRO TYR TRP GLU ASN ASN SER THR SER LEU LYS ALA LEU VAL LYS LYS PRO ASN GLY GLU LEU VAL ARG
AAG GAT TCC TCA TGT TTT TTG CCA TAT TGG GAA AAT AAT TCT ACT TCT TTA AAG GCT TTA GTG AAA AAA CCC AAT GGA GAA TTA GTT AGA

                    665
LEU THR LEU ALA THR LEU
TTA ACC TTG GCA ACA CTT TAG ATATGTAATAAAAAT GGGTGTAAACACCCA TTTTTATTTTATGTT AAATATTCTATAGCT AATTAAACCTAACA ACTATGGTTTCC

CCT ACAACACCAATATCG
```

FIGURE 2: Nucleotide sequence of gene 9 region and amino acid sequence of P22 tail protein. Numbers refer to residue positions in mature tail protein.

system.

Amino and Carboxyl Sequences of the Tail Protein. For location of the protein coding sequences in the DNA sequence, amino- and carboxyl-terminal analyses of the tail protein were performed. Purified P22 tail protein was subjected to automated Edman degradation for nine steps. The results (Table I) define the N-terminal sequence

NH$_2$-Thr-Asp-Ile-Thr-Ala-Asn-Val-Val-Val. . .

Table I: Sequential Edman Degradation of 20 nmol of Tail Protein

1	2	3	4	5	6	7	8	9
NH₂-Thr-Asp-Ile-Thr-Ala-Asn-Val-Val-Val . . .[a]								
37	91	125	23	73	46	61	64	70

Amino Acids Released by Carboxypeptidase Y Digestion[b]

	15 min	30 min	60 min	90 min
leucine	2.10	2.36	2.54	3.03
threonine	1.63	1.72	2.01	1.97
alanine	0.95	1.00	1.03	0.97
valine	0.12	0.23	0.74	0.88
arginine	0.08	0.25	0.72	0.82

[a] Numbers below residues are yields of PTH-amino acid derivatives × 10^{-10} mol. [b] Values expressed as moles of amino acid released per mole of tail protein.

Table II: Tail Protein Amino Acid Composition[a]

	time of hydrolysis (h)					
	24	48	72	96	av	seq.[e]
Asx	88.9	88.9	89.0	88.0	88.7	89
Thr	45.2	43.2	41.7	39.7	47.5[b]	46
Ser	46.0	41.3	38.3	34.0	49.8[b]	50
Glx	44.5	44.5	45.0	44.5	44.6	44
Pro	30.3	30.3	28.4	30.2	29.8	28
Gly	71.7	71.7	72.3	71.0	71.6	71
Ala	47.6	47.3	48.6	47.9	47.8	48
Val	44.9	46.9	47.7	47.6	47.6[c]	48
Met	8.5	7.3	8.5	8.2	8.1	8
Ile	46.1	47.5	49.1	48.4	48.4[c]	51
Leu	50.1	49.3	49.8	49.2	49.6	49
Tyr	27.1	27.0	27.0	26.5	26.9	27
Phe	25.1	25.0	25.3	25.0	25.1	25
His	10.2	11.9	10.7	10.4	10.8	10
Lys	33.5	33.7	33.0	35.0	33.8	34
Arg	23.1	23.1	22.2	22.7	22.7	23
Trp	ND[d]	ND	ND	ND	ND	7
Cys	ND	ND	ND	ND	ND	8

[a] Values expressed as moles of amino acid per mole of tail protein monomer. [b] Values extrapolated to zero time. [c] Values after 96 h of hydrolysis used. The partial specific volume of the tail protein calculated from its amino acid composition is 0.729 cm³/g. [d] ND, not determined. [e] Total = 666.

The DNA encoding these nine amino acids was located at bases 112–138 in the DNA sequence (Figure 2). This sequence is preceded by an ATG codon which we presume codes for the inhibitor N-formylmethionine since the prior in-frame codon is a TAA termination codon. Following the ATG initiator codon is an open reading frame sufficient to encode a protein of 666 amino acids (Figure 2). A TAG codon terminates the open reading frame at base 2110 and is immediately preceded by DNA which encodes the sequence

. . .Leu-Val-Arg-Leu-Thr-Leu-Ala-Thr-Leu-COOH

Digestion of tail protein with carboxypeptidase Y releases leucine rapidly followed by threonine, alanine, valine, and arginine (Table I). The time course and stoichiometry of amino acids released by carboxypeptidase digestion (Table I) are consistent with the C-terminal sequence inferred from the DNA sequence. We conclude that the mature gene 9 polypeptide terminates with this sequence.

The complete amino acid sequence of the tail protein deduced from the gene sequence is shown in Figure 2. The deduced and experimental amino acid compositions are compared in Table II. The agreement of these values is excellent considering the size of the tail protein. Our sequence contains eight cysteine residues compared with an experimentally determined value of ten (D. Goldenberg and P. Berget, unpub-
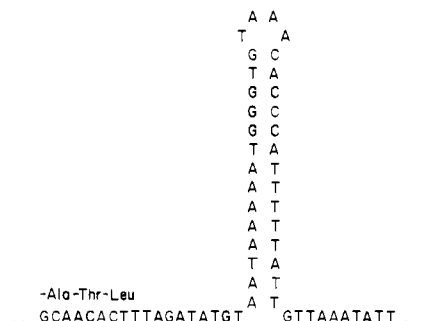

FIGURE 3: Region of hyphenated dyad symmetry, drawn as stem and loop structure, which follows C-terminal coding sequences of gene 9. Ala-Thr-Leu-COOH are the carboxyl-terminal amino acids of tail protein.

lished results). This experimental value is insensitive to prior reduction of the protein, and thus none of these cysteines seems to form disulfide bonds. The calculated molecular weight for a monomer of mature P22 tail protein is 71 759. This compares well to the M_r value of 76 000 calculated from mobility on polyacrylamide gels in the presence of sodium dodecyl sulfate (Botstein et al., 1973; Iwashita & Kanegasaki, 1976).

Discussion

We believe that the sequence reported here for the gene 9 polypeptide is highly reliable. Both the coding and noncoding strands of the gene were completely sequenced (Figure 1), and the inferred protein sequence is in excellent agreement with confirming protein chemical data (Table I). The mature tail protein polypeptide contains 666 amino acids and is predicted to be slightly acidic at neutral pH. Comparison of the gene and amino acid sequences suggests that mature tail protein arises by cleavage of the initiator N-formylmethionine from the nascent chain.

The DNA that precedes and follows the coding region of the P22 tail protein gene contains sequences characteristic of regulatory elements observed in many prokaryotic genes. Seven nucleotides prior to the ATG initiation codon, there is a sequence of DNA (5'-GGAG-3') complementary to the 3'-terminal sequence of 16S ribosomal RNA. This sequence presumably forms part of the gene 9 mRNA ribosome binding site (Shine & Dalgarno, 1974). Six nucleotides beyond the translational stop codon of the tail protein gene we observe a 32-base sequence of hyphenated dyad symmetry. In Figure 3, this sequence is shown in a stem and loop conformation. The top of the stem is GC rich and is followed by a run of T's. This sequence organization is strikingly similar to many known rho-independent transcriptional terminator sites (Platt, 1978; Steitz, 1979). We currently have no direct evidence that this sequence functions in termination of transcription, but gene 9 is the last gene in the P22 late operon, and thus it is possible that termination of late transcription occurs at this site.

No promoter-like sequences are observed in the 217 base pair region between the beginning of gene 9 and the termination codon of the adjacent antirepressor gene (R. Sauer, W. Krovatin, J. DeAnda, P. Youderian, and M. Susskind, unpublished results). Gene 9 transcription is presumed to initiate at the P22 P_late promoter some 20 kilobases upstream from gene 9 (Weinstock et al., 1980), and thus this finding is not surprising. The intercistronic region between ant and gene 9 does, however, contain a transcriptional termination site (P. Berget, A. Poteete, and R. Sauer, unpublished results). This terminator prevents tail protein transcription in the absence of a positive regulatory protein also encoded by P22 (Weinstock et al., 1980).

Several mutations that define important regions of tail protein function have been isolated and now need to be located in the primary structure. These include an endo-rhamnosidase-deficient mutant (Berget & Poteete, 1980), several monomer association defective mutants (Goldenberg & King, 1981), and assembly defective mutants (P. B. Berget and D. Raulston, unpublished results). At present, many of these mutations have been genetically mapped (P. B. Berget, G. Weinstock, and D. Botstein, unpublished results; Smith et al., 1981). A step essential for the further development of the P22 tail protein system will be a more precise correlation of the genetic and physical maps of the gene. This alignment should allow mutations to be located in the nucleotide sequence of gene 9 and thus allow the analysis of tail protein structure and function to proceed at the amino acid residue level.

References

Anderson, T. F. (1960) *Proc. Eur. Reg. Conf. Electron Microsc., 2nd*, 1008.

Backman, K. (1980) *Gene 11*, 169.

Berget, P. B., & Poteete, A. R. (1980) *J. Virol. 34*, 234.

Botstein, D., Wadell, C. H., & King, J. (1973) *J. Mol. Biol. 80*, 669.

Brauer, A. W., Margolies, M. N., & Haber, E. (1975) *Biochemistry 14*, 3029.

Chisholm, R. L., Deans, R. J., Jackson, E. N., Jackson, D. A., & Rutila, J. E. (1980) *Virology 102*, 172.

Eriksson, V., & Lindberg, A. A. (1977) *J. Gen. Virol. 34*, 207.

Geier, G. E., & Modrich, P. (1979) *J. Biol. Chem. 254*, 1408.

Goldenberg, D. (1981) Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Goldenberg, D. P., & King, J. (1981) *J. Mol. Biol. 145*, 633.

Hill, D. W., Walters, F. H., Wilson, T. D., & Stuart, J. D. (1979) *Anal. Chem. 51*, 1338.

Ilse, D., & Edman, P. (1963) *Aust. J. Chem. 16*, 411.

Israel, V., Anderson, T. F., & Levine, M. (1967) *Proc. Natl. Acad. Sci. U.S.A. 57*, 284.

Iwashita, S., & Kanegasaki, S. (1973) *Biochem. Biophys. Res. Commun. 55*, 403.

Iwashita, S., & Kanegasaki, S. (1976) *Eur. J. Biochem. 65*, 89.

King, J., Lenk, E. V., & Botstein, D. (1973) *J. Mol. Biol. 80*, 697.

Klenow, H., & Henningsen, I. (1970) *Proc. Natl. Acad. Sci. U.S.A. 65*, 168.

Maniatis, T., Jeffrey, A., & van de Sande, H. (1975) *Biochemistry 14*, 3787.

Marinus, M. G., & Morris, N. R. (1973) *J. Bacteriol. 114*, 1143.

Maxam, A. M., & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 560.

Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499.

Platt, T. (1978) in *The Operon* (Miller, J. H., & Reznikoff, W. S., Eds.) p 263, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Sauer, R. T., & Anderegg, R. (1978) *Biochemistry 17*, 1092.

Sauer, R. T., Pan, J., Hopper, P., Hehir, K., Brown, J., & Poteete, A. R. (1981) *Biochemistry 20*, 3591.

Setlow, P., & Kornberg, A. (1972) *J. Biol. Chem. 247*, 232.

Shine, J., & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. U.S.A. 71*, 1342.

Smith, D. H., Berget, P. B., & King, J. (1981) *Genetics 96*, 331.

Smith, H. O., & Birnstiel, M. L. (1976) *Nucleic Acids Res. 3*, 2387.

Steitz, J. A. (1979) in *Biological Regulation and Development I* (Goldberger, R. F., Ed.) p 349, Plenum Press, New York.

Ueda, K., McMacken, R., & Kornberg, A. (1978) *J. Biol. Chem. 253*, 261.

Weinstock, G. M. (1977) Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Weinstock, G. M., Riggs, P. D., & Botstein, D. (1980) *Virology 106*, 82.

Winston, F. (1980) Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.