Sequence of a DNA injection gene from *Salmonella typhimurium* phage P22

Pratima Adhikari and Peter B.Berget

Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh,
PA 15213, USA

Received December 7, 1992; Accepted December 11, 1992

GenBank accession no. L07556

In the Salmonella phage P22, the presence of three structural proteins in the phage particle, gp7, gp16, and gp20 is required for successful phage DNA injection (1, 2). The genes coding for these proteins are contiguous, occupy about 4 kilobases in the late operon of P22 and are located between the *imml* region and genes which specify other capsid proteins. The specific function of any one of the DNA injection gene products remains obscure. The sequence of these genes will prove useful toward the study of DNA injection in this virus. The sequences of two of these genes, gene *7* and gene *16* have recently been reported (3, 4). Here we report the sequence of gene 20.

Gene 20 is flanked on it's 5′ side by gene *7* and by gene *16* on its 3′ side. Sequencing of the gene was performed off of both strands, employing the dideoxy chain termination method (5). The sequence reveals a single 1416 base pair ORF which could code for a polypeptide of 472 amino acids with a predicted molecular weight of 50,069 daltons. The reading frame of this ORF was confirmed through the sequencing of two amber alleles, H1025 and H1032 which were identified as CAG (Gln) to TAG (amber) transition mutations (underlined in Figure 1) at codons 315 and 407 respectively. P22 particles prepared by growing the H1025 mutant under non-permissive conditions were analyzed by SDS/PAGE (not shown). These particles contain the normal complement of P22 structural proteins except for gp20 which is replaced by a polypeptide of the correct size to be the predicted 33,111 dalton amber fragment generated by this mutation. Thus as many as 158 amino acids can be missing from the carboxyl terminus of gp20 and result in a polypeptide that is still recognized for assembly into phage particles although it is not functional in DNA injection.

In the sequence that is shown in Figure 1, nucleotides 1 to 306 were previously published (4) and represent the end of gene *7* extending to nucleotide 276 of gene *20*. Nucleotides 1067 to 1454 have been previously published (3) and extend from nucleotide 1037 in gene *20* (nucleotide 1067 in Figure 1) to nucleotide 9 in gene *16*. There are three conflicts between our sequence and that of Umlauf and Dreiseikelmann (3) in the 3′ region of gene *20*. In Figure 1 the sequence GC at nucleotides 1199 and 1200 is reversed compared to their sequence; and the G residues at positions 1272 and 1402 are missing from their sequence. These differences are indicated in bold type face. We believe our sequence to be correct because this portion of the previously reported sequence was determined from only one strand (Figure 2, reference 3). There is a gap of 10 nucleotides between the stop codon of gene 7 and the start codon of gene *20* This gap contains a purine rich sequence which may be utilized as a ribosome binding site (6). The region where gene *20* ends and gene *16* begins, contains the more common stop and start codon overlap.

No significant homology was found between the gene *20* sequence and any sequence in the GenBank non-animal database. The hydropathy profile of the putative gene 20 protein was equally devoid of any notable features.

REFERENCES

1. Botstein,D., Waddell,C. and King,J. (1973) *J. Mol. Biol.* **80**, 669–695.
2. Poteete,A. and King,J. (1977) *Virology* **76**, 725–739.
3. Umlauf,B. and Dreiseikelmann,B. (1992) *Virology* **188**, 495–501.
4. Conlin,C.A., Vimr,E.R. and Miller,C.G. (1992) *J. Bacteriol.* **174**, 5869–5880.
5. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
6. Shine,J. and Dalgarno,L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1347.

```
                      20                       40
   TGCTTGGCTC GTTGTTTTAA GGGGTAATCA ATG GCT ACG TGG CAG CAG GGC ATT AAT
        GENE 7 stop >^^^            ^^^
                     GENE 20 start >Met Ala Thr Trp Gln Gln Gly Ile Asn>
    60                         80                        100
   TCA GGT GGT TTT CTG GCT GGC ATT GGT GCG CAA AAT GAG AAC GCA CCA AAG
   Ser Gly Gly Phe Leu Ala Gly Ile Gly Ala Gln Asn Glu Asn Ala Pro Lys>
               120                        140
   GCA AGA GAT ATT AAC GCA ACG CTG GGT CTG ATT CGC GAA AAC AAT GAT TTA
   Ala Arg Asp Ile Asn Ala Thr Leu Gly Leu Ile Arg Glu Asn Asn Asp Leu>
 160                       180                        200
   GCC CGT TCA GGC GCT AAT AAT GTG GCT TTA ACA GGG CTG CGT GGT CTG GCT
   Ala Arg Ser Gly Ala Asn Asn Val Ala Leu Thr Gly Leu Arg Gly Leu Ala>
             220                       240                        260
   GGC GTT GCT GAT ATT TAT AAC CAG GAA CAG CAA CAG AAA GCG CTA AAC GCA
   Gly Val Ala Asp Ile Tyr Asn Gln Glu Gln Gln Gln Lys Ala Leu Asn Ala>
                         280                       300
   TTC AAC CAG GTT CAT GCC AAC GCA TGG GCT ACT GGT GAC CCG TCT GGC CTG
   Phe Asn Gln Val His Ala Asn Ala Trp Ala Thr Gly Asp Pro Ser Gly Leu>
          320                        340                       360
   TTT AAG TTT GCT CAG GAA AAC CCT GCG TTT GTT GCG CAG GCA CAG CAG GCG
   Phe Lys Phe Ala Gln Glu Asn Pro Ala Phe Val Ala Gln Ala Gln Gln Ala>
                      380                        400
   TTT TCC GGT CTT AAT GAG CAG CAG CGT AAC GAT ATG GGC GAT TTA GCT ATG
   Phe Ser Gly Leu Asn Glu Gln Gln Arg Asn Asp Met Gly Asp Leu Ala Met>
       420                        440                        460
   AAG GCT AAC GTC GCT CTT TCT CAG GGG CCG GAA GCC TAC AGT AAA TTC ATT
   Lys Ala Asn Val Ala Leu Ser Gln Gly Pro Glu Ala Tyr Ser Lys Phe Ile>
                   480                        500
   ACT GAT AAC AAG GAC AGG TTA AAT CGT GTT GGC GCT AAT CCG GAC TGG ATG
   Thr Asp Asn Lys Asp Arg Leu Asn Arg Val Gly Ala Asn Pro Asp Trp Met>
     520                       540                        560
   ATA CAG ACT GGA GTG CAG AAT CCA GAA CAG CTA TCA CAC ATG CTG ACT ACG
   Ile Gln Thr Gly Val Gln Asn Pro Glu Gln Leu Ser His Met Leu Thr Thr>
                 580                       600
   ATG TCC CTC GGT GCG CTT GGA CCC GAA AAG GCG TTT GCT GTT CAG GAT AAG
   Met Ser Leu Gly Ala Leu Gly Pro Glu Lys Ala Phe Ala Val Gln Asp Lys>
  620                        640                       660
   ATG GTT GGC CGT GAG ATT GAC CGA GGC AGG CTG GCT GAA ACA ATC CGC AGC
   Met Val Gly Arg Glu Ile Asp Arg Gly Arg Leu Ala Glu Thr Ile Arg Ser>
              680                        700                       720
   AAT AAA GCC GGT GAG GGG CTT CAG GCT CGC GGG CAG AAT ATA ACA ATG CGT
   Asn Lys Ala Gly Glu Gly Leu Gln Ala Arg Gly Gln Asn Ile Thr Met Arg>
                          740                        760
   GGA CAA GAT ATG TCA GCG GCA ACA GCA CGA CGC GGT CAA GAT TTG GCA ACG
   Gly Gln Asp Met Ser Ala Ala Thr Ala Arg Arg Gly Gln Asp Leu Ala Thr>
           780                        800                        820
   CAA AGA GCA AAC GCC AGA ACG ATA TCA GGC AGC GAA GGA AAT CGG GTC GTT
   Gln Arg Ala Asn Ala Arg Thr Ile Ser Gly Ser Glu Gly Asn Arg Val Val>
                       840                        860
   CAG CTT GCA GAC GGG CGA ACA GTC AGC GTC GGT GGA AAA CTT CAC GGC GCA
   Gln Leu Ala Asp Gly Arg Thr Val Ser Val Gly Gly Lys Leu His Gly Ala>
         880                       900                        920
   GGG GCG AAT GCG TTT TAC GAA GGT ATT GAC GAT AAC GGC AAT ATG GTT CGT
   Gly Ala Asn Ala Phe Tyr Glu Gly Ile Asp Asp Asn Gly Asn Met Val Arg>
                     940                       960       (am H1025)T
   GTC CCG GCA AGC GCC ATT GCC GCA CCT CCA ACG TCT GCG GCA AGC GCA CAG
   Val Pro Ala Ser Ala Ile Ala Ala Pro Pro Thr Ser Ala Ala Ser Ala Gln>
      980                       1000                      1020
   AAC TAC GCA ATG AAG AAA GAC ATT GAT GCA ATC GCA AAT GCA GAT GCT TCT
   Asn Tyr Ala Met Lys Lys Asp Ile Asp Ala Ile Ala Asn Ala Asp Ala Ser>
                 1040                       1060
   GCT CTC GAT TTC ATG ACT GGA ATG ACT GGC GGA GCA GGA AAT CCG GCA ATT
   Ala Leu Asp Phe Met Thr Gly Met Thr Gly Gly Ala Gly Asn Pro Ala Ile>
  1080                       1100                       1120
   GGT GCA GAT GTT CGC AGC CGA CTC ACA GGC AAA GAG CAA CGC CAG TTA TAT
   Gly Ala Asp Val Arg Ser Arg Leu Thr Gly Lys Glu Gln Arg Gln Leu Tyr>
              1140                       1160
   AAC TCC GCA CAA CGT ATT CAG GGA AGA ATG CAG AAT CAG GGC GTG GCA GCA
   Asn Ser Ala Gln Arg Ile Gln Gly Arg Met Gln Asn Gln Gly Val Ala Ala>
1180                      1200                       1220
   GCA AGA GAT ATG GGC GCT AGC GGT ATC AAC ACC ATT GCA GAA GCG AAG ATG
   Ala Arg Asp Met Gly Ala Ser Gly Ile Asn Thr Ile Ala Glu Ala Lys Met>
            1240 (am H1032)T          1260                       1280
   TAT TTT CAG GGG ATG CCG CAG GTT GAC TAC TCA AGC CCG GAG GCT ATG CAG
   Tyr Phe Gln Gly Met Pro Gln Val Asp Tyr Ser Ser Pro Glu Ala Met Gln>
                        1300                      1320
   CAG TCT ATT CGT GAG ATT CAG GAA TAC ACC AAC AAT TAT AAC CAG CAG TAC
   Gln Ser Ile Arg Glu Ile Gln Glu Tyr Thr Asn Asn Tyr Asn Gln Gln Tyr>
         1340                       1360                      1380
   AAC GTT AAT GTT GGT AAT GGT GGG CTG AAA TCA CCA AGG CAG CAG CCA GAT
   Asn Val Asn Val Gly Asn Gly Gly Leu Lys Ser Pro Arg Gln Gln Pro Asp>
                     1400                       1420
   ACT CAG CAA TCA GCC GGA GGA AGT TAC ACG TCT AAA TCT GGC ATT AAA TTC
   Thr Gln Gln Ser Ala Gly Gly Ser Tyr Thr Ser Lys Ser Gly Ile Lys Phe>
      1440
   ACG GTG GAA TA ATG AAA GTT
  GENE 20 end >^^ ^^^
   Thr Val Glu End
   GENE 16 start >Met Lys Val
```

Figure 1. Sequence and calculated amino acid coding capacity of P22 gene 20. The locations of the two amber alleles which were sequenced are underlined. Changes associated with these mutations are indicated next to their labels. Positions of conflicts between our sequence and previously published sequence (3) are indicated by bold type face. Nucleotides for all stop and start codons in gene 20 and surrounding genes are indicated by carets below the DNA sequence and are labeled.