Volume 14 Number 18 1986 Nucleic Acids Research

The 52-protein subunit of T4 DNA topoisomerase is homologous to the gyrA-protein of gyrase

Wai Mun Huang

Department of Cellular, Viral and Molecular Biology, University of Utah Medical Center, Salt Lake City, UT 84132, USA

Received 12 August 1986; Accepted 14 August 1986

ABSTRACT

T4 gene *52* encodes one of the three subunits of T4 DNA topoisomerase. The T4 enzyme is required for normal phage DNA replication. I have cloned the entire gene, and it is expressed in uninfected *E. coli* cells. The sequence of 1966 nucleotides of T4 deletion Δsa9 surrounding gene *52* has been determined. The reading frame of the gene was established by identifying the first ten amino acids in the large open reading frame derived from the DNA sequence as those at the amino-terminus of the purified 52-protein. Based on the DNA sequence, 52-protein has 441 amino acids and a calculated peptide molecular weight of 50,583 daltons. This T4 topoisomerase subunit shares significant amino acid sequence homology with the *gyrA* subunit of bacterial gyrases and the carboxyl-half of yeast topoisomerase II in spite of the large differences in their sizes, confirming their functional equivalence in type II enzyme directed DNA topoisomerization. Amino acid sequence homology is highest in the amino-terminal portions of the equivalent peptides. The homology alignment suggests a consensus sequence organization surrounding the reactive tyrosine which is used to form the transient protein-DNA intermediate in the double-stranded DNA passing reaction.

The Δsa9 deletion in T4 brings gene *52* to a location 30 nucleotides 3' from the *rIIB* gene whose C-terminal 167 codons are also reported here.

INTRODUCTION

T4 DNA topoisomerase is a type II ATP-dependent enzyme encoded by *E. coli* bacteriophage T4. Mechanistically, it catalyzes transient double-stranded breaks in the DNA backbone through which DNA is passed resulting in changes in DNA topology. The native T4 enzyme consists of dimers of three kinds of subunits which are the products of the T4 genes *39*, *52* and *60* (1, 2). These gene products are recognized genetically as proteins required for normal T4 DNA replication and are postulated to be involved in the initiation event of T4 chromosomal DNA replication (3), although the exact roles of these proteins in phage DNA metabolism and development have yet to be clearly defined. It has been shown that T4 DNA replication in mutants defective in one of the phage DNA topoisomerase genes can be partially compensated by host gyrase (4). This result suggests that the two enzymes are interchangable in some capacity, although their *in vitro* activities are different in a fundamental way. Purified DNA gyrases are more proficient in supercoiling DNA than in relaxing DNA. The latter reaction of gyrase is ATP-independent (5). Based on the inhibitory effects of gyrase-specific antibiotics as well as other studies, it has been postulated that the

ATP-driven reaction and the concerted breakage-rejoining reaction of DNA gyrase reside on separate subunits: GyrB-protein is primarily responsible for the ATP-driven reaction and gyrA-protein is the principle subunit responsible for the breakage and rejoining activity. T4 DNA topoisomerase requires ATP for the catalytic activity of DNA relaxation and does not catalyze DNA supercoiling. In this regard, the T4 enzyme shares more similarities with the eukaryotic type II DNA topoisomerases than with DNA gyrases. Thus, it may serve as a prototype for other type II non-gyrase DNA topoisomerases which have been identified in diverse cell types (5).

The purified type II DNA topoisomerases examined thus far have different structures, although all are reported to be dimers of a protomeric structure in their native conformations (5). The T4 protomer has three peptides (60 kd, 50 kd and 18 kd); bacterial gyrases have two subunits: *gyr*A and *gyr*B; and the eukaryotic enzymes consist of one larger polypeptide. I have previously reported that the product of T4 gene *39* (coding for the 60 kd protein) is functionally and structurally equivalent to the *gyr*B subunit (6). It can be aligned with the N-terminal half of the yeast topoisomerase II protein (Huang, unpublished results).

In this communication, I report the complete nucleotide sequence of T4 gene *52*, which encodes the 50 kd subunit of T4 DNA topoisomerase. In spite of the apparent difference in enzymatic activities and sizes of the type II enzymes, significant amino acid homology can be found between 52-protein, *gyr*A subunit of gyrase and the C-terminal half of the yeast topoisomerase II. The ability to align these proteins makes it possible to identify conserved amino acids that may be critical for function.

The nucleotide sequence reported here also includes the C-terminal 167 amino acids of the T4 *rIIB* gene, which can be joined with the previously published N-terminal portion of the gene to establish the complete nucleotide sequence of T4 *rIIB* gene.

MATERIALS AND METHODS

Strains. In order to render T4 DNA digestable by general restriction enzymes, T4 quadruple mutant nd28 Δsa9 amN55 amN51 (*denA*$^-$ *denB*$^-$ *42*$^-$ *56*$^-$) (7) was used to prepare cytosine-containing T4 DNA according to a published procedure (8). The cytosine-containing DNA was digested with EcoRI and inserted into the EcoRI site of pBR322. *E. coli* 802 (9) was used as the recipient for plasmid DNA transformation. Recombinants were screened for the ability to marker rescue T4 amH17 (*52*$^-$), a mutation at the C-terminal end of the gene (10). An EcoRI fragment of about 850 bp was identified as harboring the C-terminal portion of the gene *52*. The N-terminal portion of the gene was on plasmid pN52-1 which carries a T4 HindIII-EcoRI fragment cloned into the equivalent sites of pBR313 (11). (Plasmid pN52-1 was kindly provided by T. Mattson.) The EcoRI fragment containing the C-terminus of the gene was then inserted into the EcoRI site of pN52-1 and transformed into *E. coli* B (sup°).

Recombinants were screened for their ability to complement all T4 mutants defective in gene *52*. One recombinant plasmid p1-11 was selected for further analysis. Phage M13 derivatives mp8 and mp9 were used in conjunction with *E. coli* JM101 (12) for M13 cloning and subsequent DNA sequencing analyses.

DNA sequencing. The dideoxy chain-termination method of Sanger was used according to published procedures (13, 14). α-^{35}S-dATP (NEN) was used as the label in the sequencing reactions and a 15-mer (New England Biolabs) was used as the universal primer in the synthetic reactions. The sequencing gels were dried using a slab gel drier (Bio-Rad) followed by autoradiography at room temperature with SB-5 X-ray film (Kodak).

Purification of T4 52-protein. T4 52-protein was partially purified from *E. coli* cells harboring plasmid p1-11. A 4 liter culture was grown to approximately $6x10^8$ cells/ml, and the cells were harvested by centrifugation. Cells were lysed, processed and subjected to separation by hydroxylapatite chromatography according to procedures previously described for the purification of T4 DNA topoisomerase (1). T4 52-protein was eluted from the hydroxylapatite column in the 0.3M potassium phosphate (pH 7) containing buffer. At this stage, the preparation is about 30% pure. It was further fractionated by preparative slab gel electrophoresis containing SDS (15). After staining with Coomassie brillant blue, the 52-protein band was extracted with a solution of 1% SDS to recover the 52-protein.

Other methods. The preparation of rabbit antibody against 52-protein was done using a previously published protocol (17). The procedure for Western blotting analysis and subsequent autoradiographic detection of antigen have been described (17). The amino terminal residues of 52-protein were determined using a Beckman 890C sequencer. The DNA sequence and protein comparisons were performed using the facilities provided by BIONET™. Alignments of proteins were done using the algorithm provided by Lipman and Wilbur (16) and a DEC-20 computer.

RESULTS AND DISCUSSION

The complete T4 gene 52. Plasmid p1-11 is a derivative of pBR313 carrying a 2 kb fragment of T4 DNA inserted between the HindIII and EcoRI sites (Fig 1). It was selected based on its ability to complement as well as marker rescue T4 phages defective in gene *52* in non-permissive and permissive cells, respectively (17). The alignment of the plasmid DNA with the genetic map of gene *52* immediately suggests that the gene is transcribed from the HindIII site toward the EcoRI site as described in Figure 1. Plasmid p1-11 directs the synthesis of a protein of approximately 50 kd in molecular weight, as indicated by SDS-polyacrylamide gel electrophoresis of crude extracts of the plasmid-containing cells. This is the same size as the 52-protein subunit of the T4 DNA topoisomerase. Furthermore, it was shown that this plasmid DNA encoded protein reacts with antibody prepared against T4 DNA topoisomerase. The results of such a Western-blotting analysis is shown in Fig 2a. Rabbit

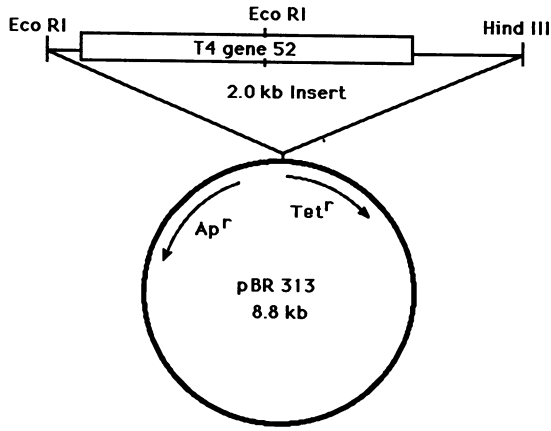

Figure 1. Restriction map of plasmid p1-11. A 2 kb region of T4 Δsa9 DNA containing gene 52 is shown as an insertion replacing the HindIII-EcoRI region of pBR313. Cells harboring the recombinant plasmid p1-11 (10.8 kb in size) is resistant to both ampicillin and tetracycline.

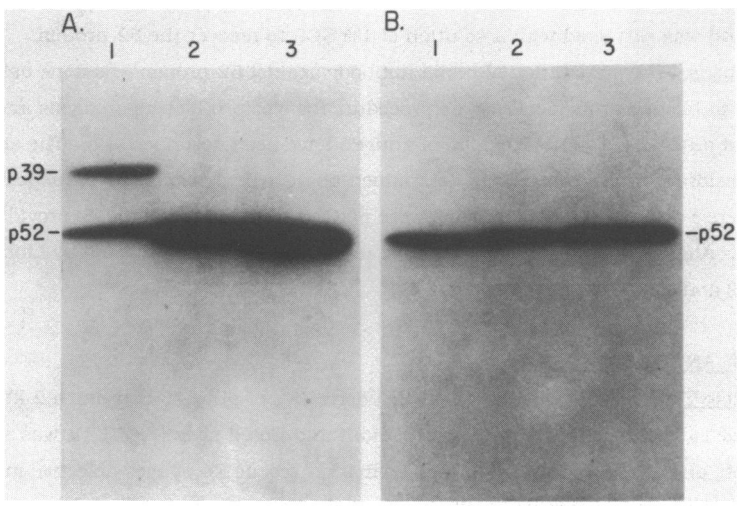

Figure 2. Western blotting analysis of T4 gene 52-protein. Purified T4 DNA topoisomerase (lane 1, 1μg) from T4 infected cells and partially purified 52-protein (lane 2, 5μg and lane 3, 10μg) from cells harboring p1-11 were analyzed on a 10% polyacrylamide gel containing SDS. The gel was blotted onto nitrocellulose membrane and allowed to interact with rabbit antibody. The subsequent radioimmune detection was performed using ^{125}I-labeled Staphylococcus protein-A. The detailed procedure has been described (17). (A). Anti-T4 DNA topoisomerase was used. (B). Anti-52-protein was used. The positions of T4 39-protein and 52-protein are indicated by p39 and p52, respectively. T4 60-protein, the third subunit in the complete T4 topoisomerase, had migrated off the gel and is not shown here.

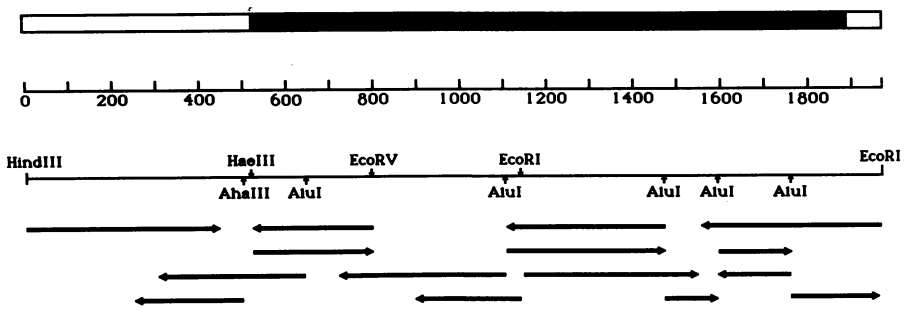

Figure 3. DNA sequencing strategy for T4 gene *52*. A restriction map of the 2 kb region is given. The fragment was digested with restriction enzymes listed in the map for subcloning and sequencing using M13-derivatives mp8 or mp9. The arrows indicate the direction and extent of the sequence determined on each fragment using the dideoxy chain-termination method. The filled region in the top bar represents the coding region of T4 gene *52*.

antibody was also prepared against the plasmid encoded 50K protein after it was purified as a denatured protein extracted from an SDS-polyacrylamide gel. Figure 2b shows that this antibody reacts specifically with the 52-protein subunit of the complete T4 DNA topoisomerase. Thus, I conclude that plasmid p1-11 habors the entire T4 gene *52*, and it is expressed in uninfected *E. coli* cells.

Sequencing strategy and nucleotide sequence. The 2 kb T4 DNA insert of p1-11 can be divided into two sections, a 850 bp EcoRI fragment and a 1140 bp HindIII-EcoRI fragment (Fig 1). The two fragments were subcloned into the appropriate sites of M13 derivatives in order to obtain template DNA for the dideoxy chain-termination method of DNA sequencing analysis. Restriction enzyme analysis revealed five AluI sites within the 2 kb fragment as shown in Figure 3. The AluI fragments were collectively cloned into the HincII site of mp9 in a shot gun fashion. The sequences from the AluI clones, many of them providing both orientations of the DNA strand, were used to provide overlapping sequences extending from areas whose sequences were obtained from the two larger fragments. Finally the unique EcoRV and AhaIII sites were chosen to generate sub-fragments for another round of cloning and sequencing in order to provide overlapping sequences within the 1140 bp HindIII-EcoRI section. The strategy is shown in Figure 3. The entire region has been covered at least twice either by sequencing a different isolate of the same sub-fragment or by sequencing overlapping fragments, and more than 75% of the region is sequenced in both directions. The nucleotide sequence is given in Figure 4. The reported sequence starts from a HindIII site and ends at an EcoRI site and contains a total of 1966 bases.

T4 gene 52-protein. Within the sequenced region, the only open reading frame large enough to accommodate a protein of approximately 50K, which plasmid p1-11 was shown to encode, starts at position 533 with an ATG initiation codon and extends to position 1855 followed by a TGA stop codon. To confirm the relationship between the open reading frame

```
                                                   34                                                   79
           A AGC TTG TTC TAT CAA AAT ATT GAA TTG CGG TCT GGT TTG GTT GAT CGT ATT CTT GAC TCG ATG GAA AAA GGC GAA AAC
             Ser Leu Phe Tyr Gln Asn Ile Glu Leu Arg Ser Gly Leu Val Asp Arg Ile Leu Asp Ser Mey Glu Lys Gly Glu Asn

                                                   124                                                  169
TTT GAA TTT TAT TTT CCG TTC TTG GAA AAT CTG TTG GAA AAC CCA AGC CAA AAA GCG GTA TCT CGA CTC TTT GAT TTC TTG GTA GCA AAC
Phe Glu Phe Tyr Phe Pro Phe Leu Glu Asn Leu Leu Glu Asn Pro Ser Gln Lys Ala Val Ser Arg Leu Phe Asp Phe Leu Val Ala Asn

                                                   214                                                  259
GAT ATT GAA ATT ACA GAA GAT GGT TAC TTC TAT GCT TGG AAA GTA GTT CGC AGC AAT TAC TTT GAC TGT CAC TCA AAC ACC TTT GAT AAC
Asp Ile Glu Ile Thr Glu Asp Gly Tyr Phe Tyr Ala Trp Lys Val Val Arg Ser Asn Tyr Phe Asp Cys His Ser Asn Thr Phe Asp Asn

                                                   304                                                  349
AGT CCG GGT AAA GTA GTT AAA ATG CCA CGT ACT CGT GTG AAT GAC GAT GAT ACA CAA ACT TGT TCT CGT GGT TTG CAT GTG TGT TCT AAA
Ser Pro Gly Lys Val Val Lys Met Pro Arg Thr Arg Val Asn Asp Asp Asp Thr Gln Thr Cys Ser Arg Gly Leu His Val Cys Ser Lys

                                                   394                                                  329
TCT TAT ATT CGT CAC TTT GGT AGT TCA ACC AGC CGA GTT GTA AAA GTT AAA GTT CAT CCT CGC GAT GTA GTA TCA ATT CCG ATT GAT TAC
Ser Tyr Ile Arg His Phe Gly Ser Ser Thr Ser Arg Val Val Lys Val Lys Val His Pro Arg Asp ''al Val Ser Ile Pro Ile Asp Tyr

                                                   484                                                  532
AAC GAT GCT AAA ATG CGT ACC TGC CAA TAT GAA GTA GTT GAA GAC GTT ACT GAA CAA TTT AAA  TAAGGCTTCG  GCCCTTAACT  AAGGAAATTT
Asn Asp Ala Lys Met Arg Thr Cys Gln Tyr Glu Val Val Glu Asp Val Thr Glu Gln Phe Lys

                                                   577                                                  622
ATG CAA CTG AAT AAT CGC GAT TTA AAA AGT ATC ATT GAT AAT GAA GCA TTG GCT TAT GCT ATG TAC ACG GTT GAA AAT CGT GCT ATC CCA
MET Gln Leu Asn Asn Arg Asp Leu Lys Ser Ile Ile Asp Asn Glu Ala Leu Ala Tyr Ala Met Tyr Thr Val Glu Asn Arg Ala Ile Pro

                                                   667                                                  712
AAT ATG ATT GAT GGA TTT AAG CCA GTT CAA CGA TTT GTT ATT GCT CGA GCT CTT GAT TTG GCA CGA GGA AAT AAA GAT AAG TTT CAC AAA
Asn Met Ile Asp Gly Phe Lys Pro Val Gln Arg Phe Val Ile Ala Arg Ala Leu Asp Leu Ala Arg Gly Asn Lys Asp Lys Phe His Lys

                                                   757                                                  802
CTC GCT TCT ATT GCA GGT GGT GTA GCG GAC CTT GGA TAT CAT CAT GGT GAA ACT CTG CAC AAG AGC CAG TGC TTG ATG GCT AAT ACT TGG
Leu Ala Ser Ile Ala Gly Gly Val Ala Asp Leu Gly Tyr His His Gly Glu Thr Leu His Lys Ser Gln Cys Leu Met Ala Asn Thr Trp

                                                   847                                                  892
AAT AAT AAC TTT CCT CTG TTA GAT GGT CAA GGA AAC TTT GGT TCT CGT ACT GTC CAA AAG GCA GCG GCA AGT CGT TAT ATT TTT GCT CGT
Asn Asn Asn Phe Pro Leu Leu Asp Gly Gln Gly Asn Phe Gly Ser Arg Thr Val Gln Lys Ala Ala Ala Ser Arg Tyr Ile Phe Ala Arg

                                                   937                                                  982
GTA AGT AAA AAT TTC TAT AAC GTA TAT AAA GAT ACT GAA TAT GCT CCG GTA CAT CAA GAT AAA GAA CAC ATT CCG CCT GCT TTC TAT TTG
Val Ser Lys Asn Phe Tyr Asn Val Tyr Lys Asp Thr Glu Tyr Ala Pro Val His Gln Asp Lys Glu His Ile Pro Pro Ala Phe Tyr Leu

                                                  1027                                                 1072
CCT ATT ATT CCT ACT GTT CTT CTT AAT GGC GTT TCC GGT ATT GCA ACT GGT TAT GCA ACT TAC ATT CTT CCT CAT AGT GTT TCT TCT GTC
Pro Ile Ile Pro Thr Val Leu Leu Asn Gly Val Ser Gly Ile Ala Thr Gly Tyr Ala Thr Tyr Ile Leu Pro His Ser Val Ser Ser Val

                                                  1117                                                 1162
AAG AAA GCT GTA CTG CAA GCT CTT CAA GGA AAG AAA GTA ACT AAA CCG AAA GTA GAA TTC CCA GAA TTT CGT GGT GAA GTC GTT GAA ATT
Lys Lys Ala Val Leu Gln Ala Leu Gln Gly Lys Lys Val Thr Lys Pro Lys Val Glu Phe Pro Glu Phe Arg Gly Glu Val Val Glu Ile

                                                  1207                                                 1252
GAT GGG CAA TAT GAA ATT CGT GGA ACA TAT AAG TTT ACT TCA CGA ACT CAA ATG CAT ATC ACT GAG ATT CCG TAT AAG TAT GAT CGT GAA
Asp Gly Gln Tyr Glu Ile Arg Gly Thr Tyr Lys Phe Thr Ser Arg Thr Gln Met His Ile Thr Glu Ile Pro Tyr Lys Tyr Asp Arg Glu

                                                  1297                                                 1342
ACT TAT GTG AGT AAA ATC TTA GAC CCA CTT GAA AAT AAA GGC TTC ATT ACA TGG GAT GAT GCT TGT GGT GAG CAT GGT TTT GGC TTC AAA
Thr Tyr Val Ser Lys Ile Leu Asp Pro Leu Glu Asn Lys Gly Phe Ile Thr Trp Asp Asp Ala Cys Gly Glu His Gly Phe Gly Phe Lys

                                                  1387                                                 1432
GTT AAA TTC CGC AAA GAA TAT TCT TTG AGC GAT AAC GAA GAA GAA CGC CAT GCA AAA ATT ATG AAA GAC TTC GGA CTG ATT GAG CGT CGT
Val Lys Phe Arg Lys Glu Tyr Ser Leu Ser Asp Asn Glu Glu Glu Arg His Ala Lys Ile Met Lys Asp Phe Gly Leu Ile Glu Arg Arg

                                                  1477                                                 1522
TCC CAG AAT ATT ACG GTT ATT AAT GAG AAA GGA AAG CTG CAA GTT TAC GAT AAC GTA GTT GAT TTA ATT AAA GAC TTT GTT GAA GTT CGT
Ser Gln Asn Ile Thr Val Ile Asn Glu Lys Gly Lys Leu Gln Val Tyr Asp Asn Val Val Asp Leu Ile Lys Asp Phe Val Glu Val Arg

                                                  1567                                                 1612
AAA ACT TAT GTC CAA AAA CGA ATT GAT AAC AAA ATT AAA GAA ACT GAG TCA GCT TTT CGT TTA GCC TTT GCC AAA GCA CAT TTC ATT AAG
Lys Thr Tyr Val Gln Lys Arg Ile Asp Asn Lys Ile Lys Glu Thr Glu Ser Ala Phe Arg Leu Ala Phe Ala Lys Ala His Phe Ile Lys

                                                  1657                                                 1702
AAA GTA ATT TCA GGT GAA ATT GTT GTA CAG GGT AAA ACT CGT AAA GAA CTG ACC GAA GAA CTT TCG AAA ATC GAT ATG TAT TCT TCT TAT
Lys Val Ile Ser Gly Glu Ile Val Val Gln Gly Lys Thr Arg Lys Glu Leu Thr Glu Glu Leu Ser Lys Ile Asp Met Tyr Ser Ser Tyr

                                                  1747                                                 1792
GTT GAT AAA CTA GTT GGA ATG AAT ATT TTT CAT ATG ACT TCC GAT GAA GCA AAG AAA CTT GCT GAA GAA GCT AAA GCT AAA AAA GAA GAA
Val Asp Lys Leu Val Gly Met Asn Ile Phe His Met Thr Ser Asp Glu Ala Lys Lys Leu Ala Glu Glu Ala Lys Ala Lys Lys Glu Glu

                                                  1837                                                 1885
AAC GAA TAT TGG AAA ACT ACT GAT GTA GTT ACT GAA TAC ACC AAA GAT TTA GAG GAA ATC AAA  TGAGTCCATT  CATTGGTATT  ACAAGCGCTG
Asn Glu Tyr Trp Lys Thr Thr Asp Val Val Thr Glu Tyr Thr Lys Asp Leu Glu Glu Ile Lys

                                                 1935
CATTAGTATC  TGGTAGCATT  TTACTGGCGG  GTTTAGGCGT  TGTTCCAGCC  GTAGCAGGAG  GTCTTCTTGC  GTTCGGAATT  C
```

Figure 4. The nucleotide sequence and the deduced amino acid sequence of the 2 kb region of T4 Δsa9 DNA. The reported sequence starts from a HindIII site near T4 map location 162.7 (29), extends for 1966 residues and ends at an EcoRI site near map location 159.6. This region contains a 2.5 kb deletion in the D region of the T4 genome. The sequenced region codes for the C-terminal 167 amino acids of T4 *rIIB* gene and the complete gene *52*. The Δsa9 deletion brings gene *52* to a location 30 nucleotides 3' to the *rIIB* gene.

Table I
Amino Acid Composition

	52-protein	rIIB-protein [a]
Ala	31	17
Arg	20	17
Asp	25	22
Asn	22	21
Cys	2	5
Glu	36	21
Gln	14	11
Gly	25	13
His	12	5
Ile	31	21
Leu	29	18
Lys	47	23
Met	9	4
Phe	21	14
Pro	13	9
Ser	21	25
Thr	25	17
Trp	3	3
Tyr	22	13
Val	33	33
Total	441	312
M.W.	50,583	35,545

a. The complete sequence of the rIIB-protein was obtained by combining the C-terminal amino acid sequence reported in Figure 4 with that previously reported for the N-terminus of the gene (25).

and the 52-protein synthesized by the plasmid, the plasmid encoded protein was isolated from an SDS-polyacrylamide gel and its N-terminal ten amino acids determined. They were found to be Met-Gln-Leu-Asn-Asn-Arg-Asp-Leu-Lys-Ser. These amino acids are in perfect agreement with the first ten amino acids derived from the open reading frame. According to the DNA sequence, T4 52-protein has 441 amino acids and a calculated peptide molecular weight of 50,583 daltons. The amino acid composition of the protein is given in Table I. Lysine is the most abundant amino acid accounting for 10.7% of the total residues. The protein has 79 basic residues and 61 acidic residues. This is also consistent with the observation that 52-protein has a basic pI value of 8.3 (Gibson and Huang, unpublished results). The frequency of codon usage in gene *52* is given in Table II. It provides a distribution similar to those calculated for the other sequenced T4 early proteins including another T4 DNA topoisomerase gene, gene *39* (6).

Homology between 52-protein and other topoisomerases. When the amino acid sequence of T4 52-protein was compared with other topoisomerases, it was found to share significant homology with the *gyr*A subunit of *B. subtilis* gyrase (18). Figure 5 shows an amino acid alignment between the two proteins. It is noted that T4 52-protein (441 amino acids), which is about half as large as the *gyr*A subunit (821 amino acids), aligns only with the N-terminal half of the gyrase protein, and no homology was recognized beyond residue number

Table II
Frequency of Codon usage in T4 Gene 52

TTT-Phe	13 (2.9)	TCT-Ser	7 (1.6)	TAT-Tyr	18 (4.1)	TGT-Cys	1 (.2)
TTC-Phe	8 (1.8)	TCC-Ser	3 (.7)	TAC-Tyr	4 (.9)	TGC-Cys	1 (.2)
TTA-Leu	6 (1.4)	TCA-Ser	3 (.7)	TAA- .	0 (.0)	TGA- .	0 (.0)
TTG-Leu	5 (1.1)	TCG-Ser	1 (.2)	TAG- .	0 (.0)	TGG-Trp	3 (.7)
CTT-Leu	9 (2.0)	CCT-Pro	5 (1.1)	CAT-His	9 (2.0)	CGT-Arg	12 (2.7)
CTC-Leu	1 (.2)	CCC-Pro	0 (.0)	CAC-His	3 (.7)	CGC-Arg	3 (.7)
CTA-Leu	1 (.2)	CCA-Pro	4 (.9)	CAA-Gln	11 (2.5)	CGA-Arg	5 (1.1)
CTG-Leu	7 (1.6)	CCG-Pro	4 (.9)	CAG-Gln	3 (.7)	CGG-Arg	0 (.0)
ATT-Ile	25 (5.7)	ACT-Thr	19 (4.3)	AAT-Asn	15 (3.4)	AGT-Ser	5 (1.1)
ATC-Ile	6 (1.4)	ACC-Thr	2 (.5)	AAC-Asn	7 (1.6)	AGC-Ser	2 (.5)
ATA-Ile	0 (.0)	ACA-Thr	2 (.5)	AAA-Lys	36 (8.2)	AGA-Arg	0 (.0)
ATG-MET	9 (2.0)	ACG-Thr	2 (.5)	AAG-Lys	11 (2.5)	AGG-Arg	0 (.0)
GTT-Val	17 (3.9)	GCT-Ala	17 (3.9)	GAT-Asp	21 (4.8)	GGT-Gly	12 (2.7)
GTC-Val	4 (.9)	GCC-Ala	2 (.5)	GAC-Asp	4 (.9)	GGC-Gly	3 (.7)
GTA-Val	11 (2.5)	GCA-Ala	10 (2.3)	GAA-Glu	30 (6.8)	GGA-Gly	9 (2.0)
GTG-Val	1 (.2)	GCG-Ala	2 (.5)	GAG-Glu	6 (1.4)	GGG-Gly	1 (.2)

468 of the *B. sublitis* protein. The overall homology is 31.7% of identical amino acids at corresponding positions. With the computer generated alignment, it is also noted that more conserved amino acids are clustered in the first 200 residues and more and larger gaps have to be included in the remaining regions in order to obtain optimal alignment.

T4 52-protein also shares significant homology with the C-terminal end of yeast DNA topoisomerase II (19). It has recently been shown that this portion of the yeast protein, which consists of 763 residues, aligns with the *gyr*A subunit of the *B. subtilis* gyrase (20). An amino acid alignment between 52-protein and yeast topoisomerase II is given in Figure 6. It shows that homology is displayed in two blocks with a loop of 246 residues present in the larger yeast protein, thereby accommodating the difference in size between the proteins. The overall homology is about 29%. This value is comparable to the homology recognized in the comparison between T4 52-protein and *B. subtilis gyr*A protein (Fig 5) and is higher than the homology recognized between yeast topo II and gyrase (20). It is also noted that, similar to the alignment with the gyrase subunit, more conserved amino acids are present at positions in the first 200 amino acids of the 52-protein, and many of the conserved amino acids are present in all three proteins at corresponding positions. The most striking homology is shown in a stretch of about 75 amino acids from 101 to 175 in the 52-protein coordinates as given in Figure 7. The DNA sequence of *E. coli gyr*A subunit has recently been determined (Wang, unpublished results). It shares 54.8% overall homology with the corresponding *B. subtilis* protein. Yet the specific regions examined here (from amino acids 108 to 186 in the *B. subtilis* protein coordinates) are nearly identical between the two

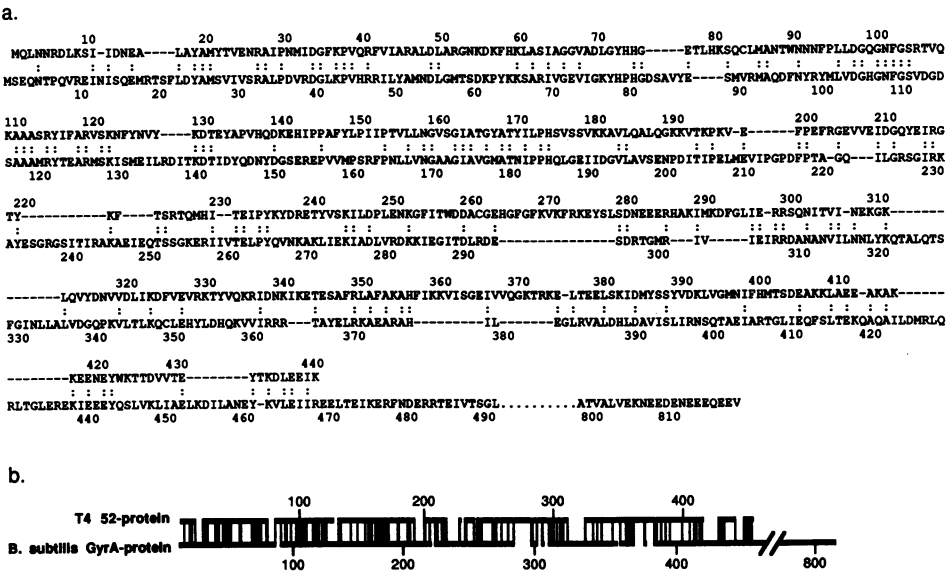

Figure 5. Amino acid homology between T4 gene 52-protein and the gyrA-protein of *B. subtilis* gyrase. (a) Amino acid alignment between the two proteins. T4 52-protein is given in the top line and the *B. subtilis* protein is given in the bottom line. Homology is mainly located in the N-terminal half of the gyrA-protein, hence gyrA amino acids between residues 493 to 799 (18) are omitted as indicated by a dotted line. (b) A schematic drawing of the homology described above. Pairs of identical amino acids are represented by vertical lines drawn between the linear sequences. Breaks in the sequences represent amino acid gaps which are included in order to maintain optimal homology.

proteins. The pertinent region of the *E. coli* protein is also included in Fig 7 for comparison.

One of the hallmarks of a DNA topoisomerase reaction is the ability of the enzyme to form a covalent complex with its substrate DNA. When a type II DNA topoisomerase reaction is interrupted by the addition of a protein denaturant, the superhelical DNA substrate is cleaved into linear duplex molecules with one of the topoisomerase subunits becoming covalently attached to the 5'-ends of the DNA via a phospho-tyrosyl linkage (5). The attached subunit is thought to be involved with the concerted DNA cutting and rejoining portion of the complete enzymatic reaction. In the DNA gyrase reaction, it is the *gyr*A subunit which is attached to the cleaved DNA in the induced cleavage reaction (5, 21). Correspondingly, it has been suggested that the 52-protein is attached to the cleaved DNA in a T4 DNA topoisomerase directed reaction (22; Huang, unpublished results). Therefore, it is likely that the 52-protein and the *gyr*A subunit may play a similar role in the respective DNA topoisomerase reactions. Recently, it has been demonstrated with *E. coli* DNA gyrase that the DNA-protein complex generated by the addition of protein denaturant is via Tyr-122 of *gyr*A (20). *E. coli* Tyr-122 of *gyr*A subunit is situated within the highly conserved region shared by these type II topoisomerase subunits (Fig 7). By this analogy, I would propose that

a.

```
           10                  20         30        40        50        60        70        80        90        100
MQLNNRDLKSIIDNEA---------LAYAMYTVENRAIPNMIDGFKPVQRFVIARALDLARGNKDKFHKLASIAGGVADLGYHHGETLHKSQCLMANTWN---NNFPLLDGQGNFGSRTV
       :: :              ::          :::  ::::: :: :             :   ::          :::::                 ::  ::   : :: :
————DPTLKEIPISDFINKELILFSLADNIRS-----IPNVLDGFKPGQRKVLYGCFKKNLKSELKVAQLAPYVSECTA--YHHGEQSLAQTIIGLAQNFVGSNNIYLLLPNGAFGTRAT
         670       680       690           700       710       720       730           740       750       760       770

110          120        130       140                        150       160       170       180       190       200
QK--AAASRYIFARVSKNFYNVYKDTEYAPVHQDKEHIPPAF------------------YLPIIPTVLLNGVSGIATGYATYILPHSVSSVKKAVLQALQGKKVTKPKVEFPEFRGEVV
    ::: :::              ::         :  ::                   :::: :  : ::  :: ::  ::: :       :                     :::
GGKDAAAARYIY-------------TELNKLTRKIFH--PADDPLYKYIQEDEKTVEPEWYLPILPMILVNGAEGIGTGRSTYIPPFNPLEIIKNIRHLMNDEELEQMHPW---FRGWTG
         780                 790       800       810       820       830       840       850       860             870

210       220                230       240                  250       260       270       280       290       300
EIDGQYEIRGTYKFTSRTQMH-------ITEIPYKYDRETYVS----------KILDPLENKGFITWDDACGEHGFGFKVKFRKEYSLSDNEEERHAKIMKDFG--LIERRSQNITVINE
 :    ::                    ::: :      :             ::            ::           ::     ::  :     ::       ::   :    :
TIE---EIEPLRYRMYGRIEQIGDNVLEITELPARTWTSTIKEYLLLGLSGNDKIKPWIKDMEEQH-DDNI---------KFIIT--LSPEEMAKTRKIGFYERFKLISPISLMNMVAFD
         880       890       900       910       920       930       940                   950       960       970

310       320       330       340       350       360       370                   380       390       400
K-GKLQVYDNVVDLIKDFVEVRKTYVQKRIDNKIKETESAFRLAFAKAHFIKKVISGEIVVQGKTRKELTE---------------ELSKIDMYSSYVDKLVGMNIFHM------TSDEA
  ::   :  :      :  ::  :  ::: :                  :::  :           ::::                :::::              ::        :: :
PHGKIKKYNSVNEILSEFYYVRLEYYQKRKDHMSERLQWEVEKYSFQVKFIKMIIE----------KELTV——————//———————ELSKISNKFKKIST-----IFDKMGSTSATSKEN
         990      1000      1010      1020      1030          1040                      1290          1300      1310

            410                  420       430       440
---------KKLAEEAK--------AKKEENEYWKTTDVVTEYTKDLEEIK
         ::    ::        :::      :   ::        ::
TPEQDDVATKKNQTTAKKTAVKPKLAKKPVRKQQK---VVELSGESDLEILDSY——//——GSDVSFNEED
       1320      1330      1340      1350      1360         1420
```

b.

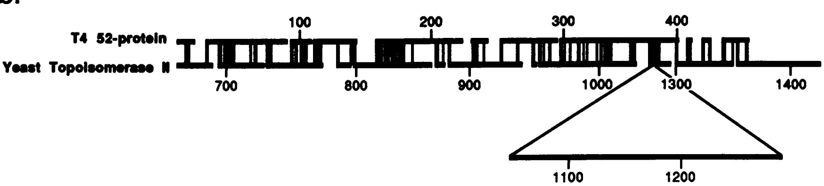

Figure 6. Amino acid homology between T4 gene 52-protein and yeast DNA topoisomerase II. (a) Amino acid alignment between the two proteins. T4 protein is given in the top line and the yeast protein (19), starting from residue 665, is given in the bottom line. Thick broken lines represent large blocks of amino acids in the yeast protein which share no homology with the T4 protein and are therefore omitted for the sake of simplicity. (b) A schematic drawing of the homology described above. Pairs of identical amino acids are indicated by vertical lines joining the horizontal linear sequences. Breaks are included to optimize the alignment. The projection from the yeast protein represents the loop region which shares no homology with the T4 protein.

Tyr-116 of T4 52-protein is involved in the reactive site of T4 DNA topoisomerase for the cutting and rejoining of DNA. Although the single peptide yeast enzyme, the three peptide T4 enzyme and the two peptide gyrases have different characteristics, the homologies shown

```
yeast        766 -G-A-F-G-T-(7)-A-A-A-A-R-Y-I-(39)-P-M-I-L-V-N-G-A-E-G-I-G-T-G-R-S-T-Y-I-P-P-F-  845
                 *   * *       * * *   * * :      *     *   * *     * *   : *     * : *   *
T4           101 -G-N-F-G-S-(5)-A-A-A-S-R-Y-I-(36)-P-T-V-L-L-N-G-V-S-G-I-A-T-G-Y-A-T-Y-I-L-P-H-  175
                 * : * * :     * * *   * *        *     *   * *     * * :   *   : *   *   * :
B. subtilis  108 -G-N-F-G-S-(5)-A-A-A-M-R-Y-T-(40)-P-N-L-L-V-N-G-A-A-G-I-A-V-G-M-A-T-N-I-P-P-H-  186
                 * : * * :     * * *   * *        *     *   * *     * * :   *   : *   *   * :
E. coli      107 -G-N-F-G-S-(5)-A-A-A-M-R-Y-T-(40)-P-N-L-L-V-N-G-S-S-G-I-A-V-G-M-A-T-N-I-P-P-H-  185
                                         |
                                   reactive site
                                         |
                                         |
Consensus        -G-X-F-G-(6-8)-A-A-A-X-R-Y-(37-41)-P-X-X-L-X-N-G-X-X-G-I-X-X-G-X-X-T-X-I-X-P-
```

Figure 7. Amino acid sequences surrounding the tyrosine reactive sites in DNA topoisomerases. * indicates conserved amino acids found in all four proteins, : indicates amino acids common among the prokaryotic proteins (T4, *B. subtilis* and *E. coli*) or between the non-gyrase enzymes (T4 and yeast). In the consensus sequence, the conserved amino acids are represented in bold one letter code, and X represents the variable amino acids.

in Figure 7 also suggest that all three kinds of type II DNA topoisomerases may use a similar strategy and local structural design to accomplish the cutting and rejoining function of the topoisomerization reaction. A consensus sequence can be generated to describe the flanking regions of the reactive site (Fig 7). Figure 7 also shows that in addition to the well-conserved amino acids, there are residues common only to the T4 and yeast proteins and some are common only among the prokaryotic proteins. This group of residues may represent the divergence between prokaryotes and eukaryotes. It is also possible that some of the former residues may also play a role in modulating the DNA strand-passage reaction in such a way that the linking number of DNA can only be increased, as in the cases of T4 and yeast enzymes. They are capable of untwisting but not twisting superhelical DNA; and gyrases are capable of both increasing and decreasing the linking number of DNA.

The homologies between the C-terminal end of 52-protein and the respective regions of the yeast protein and the *gyr*A subunit (Figs 5 and 6) are more difficult to rationalize at present. Although significant homology can be recognized in the pairwise comparison, the homologies are not conserved among different pairs. The fact that T4 52-protein has only 441 amino acids but apparently contains sufficient information to perform the cutting and rejoining function of the complete DNA topoisomerase also implies that the C-terminal half of the *gyr*A subunit, and the large loop in the yeast protein may contain system specific information relating to the involvement of these topoisomerases in more complex cellular metabolism not found in the T4 system.

The sequence of the C-terminal portion of T4 *rIIB* gene. The T4 DNA used in this report is derived from T4 Δsa9 which contains a 2.5 kb deletion in the non-essential D-region located between T4 genes *rIIB* and *52* (23). The DNA sequence reported in Figure 3 shows another open reading frame containing 167 amino acids starting from the beginning HindIII site. I therefore determined if the C-terminal portion of the *rIIB* gene is located in plasmid p1-11. Plasmid p1-11 specifically marker rescues *rIIB* C-terminal mutants N29 and 1519 (24). Since it has been reported that there is only one HindIII site within the *rIIB* gene (Shinedling and Gold, unpublished results), the complete *rIIB* gene can be obtained by joining the sequence coding for the 167 amino acids in frame with the published N-terminal sequence of the gene which ends with a HindIII site (25). Thus, the entire rIIB gene encodes a protein of 312 amino acids with a calculated peptide molecular weight of 35,545 daltons. This value is in good agreement with that determined for the rIIB-protein from T4 infected cells (15). The amino acid composition of rIIB-protein is given in Table I. The T4 *rIIB* gene is one of the best studied loci of phage T4, and its genetic analysis has yielded many of the the classical principles of molecular biology. The availability of the complete sequence of the gene will facilitate a more precise understanding of the various phenotypes associated with the large collection of T4 *rII* mutants. In spite of the extensive genetic information, the function of the rIIB-protein in T4 physiology remains undefined. It is

generally thought that the protein is readily associated with the bacterial membrane (26, 27). The DNA sequence established in this report shows that the protein has as its most abundant amino acid the hydrophobic amino acid valine (10.6%). Furthermore, it displays two hydrophobic regions between residues 55 and 95 with a maximun average hydropathy value of about 2.54, according to the rules of Kyte and Doolittle (28). Whether these hydrophobic regions are responsible for its observed membrane associability remains to be proven experimentally. The available sequence of the gene will make rIIB-protein more amenable to further biochemical analysis.

ACKNOWLEDGEMENTS

I thank S. R. Casjens for reading of the manuscript and James C. Wang for communication of results prior to publication. This work was supported by National Institutes of Helath Grant GM-21960.

REFERENCES

1. Stetler, G., King, G. and Huang, W. M. (1979) Proc. Natl. Acad. Sci. USA, 76, 3737-3742.
2. Liu, L., Liu, C. and Alberts, B. (1979) Nature, 281, 465-461.
3. Kreuzer, K. and Huang, W. M. (1983) in Bacteriophage T4, edited by Mathews, K., Kutter, E., Mosig, G. and Berget, P., A. S. M. Publications, p. 90-96.
4. McCarthy, D. (1979) J. Mol. Biol., 127, 265-283.
5. Wang, J. (1985) Ann. Rev. Biochem., 54, 665-697.
6. Huang, W. M., Nucleic Acids Res., manuscript summitted.
7. Mattson, T., Van Houwe, G., Bolle, A., Selzer, G., Epstein, R. (1977) Molec. gen. Genet., 154, 319-326.
8. Synder, L., Gold, L. and Kutter, E. (1976) Proc. Natl. Acad. Sci. USA, 73, 3098-3102.
9. Selzer, G., Bolle, A., Krisch, H. and Epstein, R. (1978) Molec. gen. Genet., 159, 301-309.
10. Mufti, S. and Bernstein, H. (1974) J. Virology, 14, 860-871.
11. Bolivar, F., Rodriguez, R., Betlach, M. and Boyer, H. (1977) Gene, 2, 75-93.
12. Messing, J. and Vieira, J. (1982) Gene, 19, 262-276.
13. Sanger, F., Nicklen, S. and Coulson, A. (1977) Proc. Natl. Acad. Sci. USA, 74, 5463-5467.
14. Higgin, M., Gibson, T. and Hong, G. (1983) Proc. Natl. Acad. Sci. USA, 80, 3963-3965.
15. O'Farrell, P., Gold, L. and Huang, W. M. (1973) J. Biol. Chem., 248, 5499-5501.
16. Wilbur, W. and Lipman, D. (1983) Proc. Natl. Acad. Sci. USA, 80, 726-730.
17. Huang, W. M., Wei, L. and Casjens, S. (1985) J. Biol. Chem., 260, 8973-8977.
18. Moriya, S., Ogasawara, N. and Yoshikawa, H. (1985) Nucleic Acids Res., 13, 2251-2265.
19. Glaever, G. Lynn, R., Goto, T. and Wang, J. C. (1986) J. Biol. Chem., in the press.
20. Lynn, R., Glaever, G., Swanberg, S. and Wang, J. C. (1986) Science, in the press.
21. Sugino, A., Higgins, P. and Cozzarelli, N. (1980) Nucleic Acids Res., 8, 3865-3874.
22. Rowe, T., Tewey, K. and Liu, L. (1984) J. Biol. Chem., 259, 9177-9181.
23. Depew, R., Snopek, T. and Cozzarelli, N. (1975) Virology, 64, 144-152.
24. Parma, D., Dill, M. and Slocum, M. (1979) Genetics, 92, 711-720.
25. Pribnow, D., Sigurdson, D., Gold, L., Singer, B., Napoli, J., Brosius, T., Dull, T. and Noller, H. (1981) J. Mol. Biol., 149, 337-376.
26. Huang, W. M. (1975) Virology, 66, 508-521.
27. Takac, B. and Rosenbusch, J. (1975) J. Biol. Chem., 250, 2339-2350.
28. Kyte, J. and Doolittle, R. (1982) J. Mol. Biol., 157, 105-132.
29. Kutter, E. and Ruger, W. (1983) in Bacteriophage T4, edited by Mathews, K., Kutter, E., Mosig, G. and Berget, P., A. S. M. Publications, p. 277-290.