Isolation, Characterization, and Expression in *Escherichia coli* of the DNA Polymerase Gene from *Thermus aquaticus*

(Received for publication, August 31, 1988)

Frances C. Lawyer, Susanne Stoffel, Randall K. Saiki, Kenneth Myambo, Robert Drummond, and David H. Gelfand

From the Departments of Microbial Genetics, Human Genetics, and Protein Chemistry, Research Division, Cetus Corporation, Emeryville, California 94608

The thermostable properties of the DNA polymerase activity from *Thermus aquaticus* (Taq) have contributed greatly to the yield, specificity, automation, and utility of the polymerase chain reaction method for amplifying DNA. We report the cloning and expression of Taq DNA polymerase in *Escherichia coli*. From a λgt 11:Taq library we identified a Taq DNA fragment encoding an epitope of Taq DNA polymerase via antibody probing. The fusion protein from the λgt 11:Taq candidate selected an antibody from an anti-Taq polymerase polyclonal antiserum which reacted with Taq polymerase on Western blots. We used the λgt 11 clone to identify Taq polymerase clones from a λCh 35:Taq library.

The complete Taq DNA polymerase gene has 2499 base pairs. From the predicted 832-amino acid sequence of the Taq DNA polymerase gene, Taq DNA polymerase has significant similarity to *E. coli* DNA polymerase I. We subcloned and expressed appropriate portions of the insert from a λCh 35: polymerase under control of the lac promoter.

Taq DNA polymerase (Taq Pol I) isolated from *Thermus aquaticus* has been shown to be highly useful in the polymerase chain reaction (PCR) method (1, 2) of amplifying DNA fragments (3). The high temperature optimum activity, 75 °C, affords unique advantages when comparing Taq Pol I to *Escherichia coli* DNA polymerase I. High specificity of primer binding at the elevated temperature gives a higher yield of the desired product with less nonspecific amplification product. Also, *E. coli* DNA polymerase I is inactivated at 93–95 °C, the temperature range required to denature the duplex DNA product. Since Taq Pol I is stable at 93–95 °C, one can add Taq Pol I only at the beginning of the PCR reaction rather than before each round of amplification.

A 62–63-kDa Taq Pol I has been purified from *T. aquaticus*, but growing the organism is more difficult than *E. coli* and polymerase yields are low (4, 5). We have developed an alternative purification protocol yielding a 94-kDa enzyme with 10–20 times higher specific activity than that previously reported. While the activity yield is quite high (40–60%), the initial expression level of Taq DNA polymerase in the native host is quite low (0.01–0.02% of total protein). Therefore, we sought to clone the Taq Pol I gene and express the gene in *E. coli*. In addition, the availability of the enzyme and the DNA sequence of the Taq DNA polymerase gene will facilitate the study of structure/function relationships and permit detailed comparisons with mesophilic DNA polymerases.  

**MATERIALS AND METHODS**

**RESULTS**

λgt 11 Libraries—The construction of three λgt 11:Taq libraries is described under "Materials and Methods," in the Miniprint. To maximize the probability of recovering a Taq Pol I epitope, three separate AluI libraries were prepared. We ligated 8-mer, 10-mer, and 12-mer EcoRI linkers to the Taq AluI DNA fragments to ensure that each AluI fragment would be in-frame with respect to β-galactosidase in one of the libraries. Upon screening with primary antibody from Taq Pol I-immunized rabbits and plaque purification, we identified seven positive plaques from the 12-mer library, four positive plaques from the 10-mer library, and no positive plaques from the 8-mer library. The EcoRI inserts fell into four size classes: two of the seven phage isolated from the 12-mer library and two of the four phage isolated from the 10-mer library contained 115-bp inserts; five clones from the 12-mer library had inserts of 175 bp (one of these also had a second apparently unrelated EcoRI fragment of 185 bp), one clone from the 10-mer library had a 125-bp insert, and one clone from the 10-mer library had a 160-bp insert. Upon antibody screening each of the phage reacted with immune serum but did not react with preimmune serum. ³²P-labeled probes were prepared by PCR amplification (3) of one clone each from the 115-, 175-, and 125-bp size classes. The 115-bp probe hybridized with all the candidates containing 115-bp inserts and no others. Similarly, the 175-bp probe hybridized with candidates containing 175-bp inserts, and the 125-bp probe hybridized...
with only the candidate containing that insert. Subsequent DNA sequencing of two 115-bp EcoRI inserts, one each from the 12-mer and 10-mer libraries, confirmed that they were identical sequences. DNA sequence analysis of Taq and flanking lacZ DNA for the candidate from the 12-mer library indicated the presence of one EcoRI linker at its 5' lacZ junction. DNA sequence analysis of the Taq and flanking lacZ DNA for the candidate from the 12-mer library indicated the presence of three EcoRI linkers at the 5' lacZ junction, which resulted in the same frame with respect to galactosidase as that of the 12-mer linker candidate. Thus, we picked DNA fragments encoding the same epitope from two libraries.

Lysogens were made of all the candidates in strain Y1089 and were induced with isopropyl-1-thio-β-D-galactopyranoside (IPTG). Total proteins from crude lysates of induced cultures were run on SDS-PAGE gels, and Western blots were prepared by using the anti-Taq Pol I antibody for detection. All of the clones made IPTG-inducible, lacZ-fusion proteins which reacted with the anti-Taq Pol I antibody (data not shown).

One clone each from the 115-, 125-, 160-, and 175-bp insert size classes was chosen for epitope selection. This method uses crude extracts of candidate clones to select antibodies from a polyclonal antiserum. These affinity-selected antibodies were used to probe Western blots of Taq Pol I. The results are shown in Fig. 1. In two experiments candidate λgt11 1, the 115-bp insert candidate, was the only one of the four tested which successfully bound antibody that reacted with purified Taq Pol I and reacted uniquely with Taq Pol I in crude extracts. The other three candidates, which had been identified and purified with the anti-Taq Pol I antibody, failed to “fish” from that same polyclonal antibody an antibody that would react with Taq Pol I on a Western blot. A close inspection of the Western blot indicates a faint cross-reaction with 28-30-kDa proteins in total soluble Thermus crude extracts. The DNA sequences of these three candidates do not correspond to any part of the Taq Pol I DNA sequence (Fig. 2).

**Ach35 Libraries**—The 115-bp EcoRI fragment from clone λgt11 1 was subcloned into Genescribe Z vector pTZ19R to use as a probe in screening the Ach35:Taq library. Construction of the partial Sau3A digest library of Taq DNA in Ach35 and screening of the library are detailed under “Materials and Methods,” in the Miniprint. The in vitro packaged library was plated initially on E. coli strain K802. That strain was chosen to avoid the possibility of degradation of Tq insert DNA by the mcrA or mcrB restriction systems (6). The amplified library was subsequently plated on E. coli strain MC1000.

Nine candidates were isolated and purified from the Ach35:Taq library. From restriction analysis of mini DNA preparations, none of the candidates proved to be identical, though they all shared some common restriction fragments. Upon Southern blotting, the pTZ19R-1 probe hybridized to a common 4.2-kb BamHI fragment and a common 6.5-kb PstI fragment in all the candidates, consistent with the hybridization seen in Southern blots of Taq genomic DNA (Fig. 3). For HindIII, the probe hybridized to fragments of different sizes, ranging in size from 5.6 to 10 kb. In addition, all nine candidates shared a common 4.5-kb HindIII fragment.

One candidate, designated d4-2, had a probe-hybridizing HindIII fragment of approximately 8 kb which corresponded to the HindIII fragment that hybridized with probe 1 in the Taq genomic Southern (Fig. 3). We chose this candidate for further study and subcloned each of its four detectable HindIII fragments (A = 8 kb, B = 4.5 kb, C = 0.8 kb, and D = 0.5 kb) into vector BSM13" in both orientations, transforming into host DG98. The two subclones of fragment A in both orientations, pFC82.35 and pFC82.2, were IPTG-induced and extracts were assayed for Taq Pol I activity (Table I). Subclone pFC82.35 had IPTG-inducible thermostable activity at a very low level, which was detectable because of the high sensitivity of the assay (<1 molecule/10 cell equivalents). In contrast, pFC82.2 had a significantly lower basal level of Taq Pol I activity which was attenuated in extracts of IPTG-grown cultures.

A restriction map of the A fragment was generated and is shown in Fig. 4. Southern analysis showed that the λgt11 1 probe hybridized at one end of the A fragment. Indeed, the DNA sequence of the A fragment isolated in λgt11 1 corresponds to nucleotides 619-720 in the Taq Pol I gene (Fig. 2). Further, the EcoRI-adapted AluI site at the junction between E. coli lacZ and Taq in λgt11 1 corresponds to the lac promoter-proximal Taq HindIII site in pFC82.35.

**Deletions in the A Fragment to Localize the Taq Pol Gene**—Two different deletions were made in the A fragment in pFC82.35 to aid in localizing the gene. In pFC84, approximately 2.4 kb of the right end of the A fragment was deleted from the SpHl site (Fig. 4) rightward to the SpHl site in the vector polylinker. In pFC85, approximately 5.2 kb of the right end of the A fragment was deleted from the Asp718 site rightward (Fig. 4) to the Asp718 site in the vector polylinker, leaving 2.8 kb of Taq insert sequence. The activity of Taq Pol I was assayed in extracts of uninduced and IPTG-induced pFC84 and pFC85 in DG101. As can be seen in Table I, deleting 3'C sequences in the A fragment had a dramatic effect on the IPTG-inducible expression of Taq Pol I. In addition, while we were unable to detect Taq Pol I in Western blots of IPTG-induced pFC82.35/DG98, induced immunoreactive bands were clearly seen upon Western blotting of IPTG-induced pFC84/DG101 and pFC85/DG101. Lanes resolved doublet immunoreactive bands that were approximately 65- and 68-kDa. These immunoreactive species were considerably smaller than full-length 94-kDa Taq Pol I. We determined that the doublet bands were not artifacts of the gel analysis because they were seen repeatedly in several experiments.

**lacZa Fusions**—To define further the locus of the Taq Pol I gene and to confirm the reading frame at different sites for use as guideposts during DNA sequence analysis, we constructed several fusions of the left end of the Taq HindIII A fragment to lacZa in the BSM13” vector. These fusions are...
Isolation and Expression of Taq Pol I Gene

Nucleotides were numbered consecutively from the start of the gene. Nucleotide numbers are shown on the left. Amino acid numbers are shown on the right.
amplified probe contains the Xgtll primer sequences on either end with $^{32}$P-labeled PCR-amplified probe. Lane 1 is a size standard EcoR1- and BamHI-digested $\lambda$pac5 andMsp1-digested plasmid Lac5. DNA fragment sizes (in kilobases) are listed at left. The PCR-amplified probe contains the Xgtll primer sequences on either end (flanking the EcoRI site in lacZ) which are homologous to sequences in the 14,300 and 6,700 marker bands. Lanes 2-6 are Taq genomic DNA digested with HindIII, HindIII and PstI, PstI, and BamHI, respectively.

DNA digested with HindIII, HindIII and PstI, PstI, and BamHI, respectively.

Lanes 2-6 are Taq genomic DNA digested with HindIII, HindIII and PstI, PstI, and BamHI, respectively.

**TABLE I**

| Experiment | Extract | IPTG | Specific activity* |
|------------|---------|------|--------------------|
| I          | BSM13*  | ±    | <0.001            |
|            | BSM13* w/Taq* | + | 0.142             |
|            | BSM13* w/Taq* | + | 0.136             |
|            | pFC82.35 | -   | 0.248             |
|            | pFC82.2 | -    | 0.301             |
|            | pFC82.2 | +    | 0.031             |
|            | pFC82.2 | +    | 0.002             |
| II         | BSM13*  | +    | 0.005*            |
|            | pFC84   | -    | 1.24              |
|            | pFC84   | +    | 29.7              |
|            | pFC85   | -    | 0.87              |
|            | pFC85   | +    | 29.6              |
|            | pLSG1   | 4.4  | 37.5              |

* Specific activity in units/mg total crude extract protein when assayed, as described under "Materials and Methods," on clarified, heat-treated extracts.

A background of 0.004% input counts has been subtracted. Extract protein corresponding to $3 \times 10^7$ cells was assayed.

* Purified Taq DNA polymerase was added to a replicate cell pellet at time of lysis. The assay contained $4 \times 10^7$ molecules of Taq Pol I.

* Purified Taq Pol I, corresponding to $4 \times 10^7$ molecules, was admixed with the BSM13* extract at time of assay.

* A background of 0.002% input counts has been subtracted. BSM13* specific activity represents two times background.

Assembled the Full-length Taq Pol I Gene—As described above, the Sphl and Asp718 deletants, pFC84 and pFC85, produced thermostable polymerase activity upon induction. However, the size of the induced bands detected by anti-Taq Pol I antibody in Western blots was smaller than full-length Taq Pol I, i.e. approximately 65-kDa as opposed to full-length 94-kDa. Thus, we felt that the A fragment lacked the 5' portion of the gene which would encode the N terminus.

Also mentioned earlier, all candidates from the $\lambda$Ch35 library which had been identified with the pTZ19R 1 probe shared a common, approximately 4.5-kb HindIII fragment which did not hybridize to the probe. This fragment, the B fragment, was subcloned into BSM13*, yielding plasmid pFC83. The restriction map of the B fragment was determined (Fig. 4). By comparing those mapping results and the A fragment map with the results of Taq genomic Southern blots probed with probe 1 (Fig. 3) we deduced that HindIII fragment B was likely to contain the 5' portion of the Taq Pol I gene.

The 724-bp BglII-HindIII segment of the B fragment was subcloned into BamHI- and HindIII-digested BSM13*. Upon sequencing, an ATG and subsequent open reading frame was found 109 bp downstream of the BglII site. The open reading frame continued to the HindIII site. In addition, the phase of the open reading frame at the "right" end of the B fragment was identical to the phase of the open reading frame at the "left" end of the A fragment.

PCR amplification confirmed that the B and A fragments in pFC83 and pFC82.35 are contiguous in the Taq genome. Primers were chosen which flanked the presumed internal HindIII site: MK138 (Table V, in the Miniprint) on the left side of HindIII and FL25, a 20-mer complementary to nucleotides 622-641 of the Taq Pol I sequence, on the right side of HindIII. Upon amplification (3) of the $\lambda$Ch35 genomic phage...
The DNA sequence was determined at the site of each fusion.

The lacZa phenotype was determined on agar plates containing X-Gal. In-frame fusions resulted in blue colonies on X-Gal and out-of-frame fusions yielded white colonies.

The DNA sequence was determined at the site of each fusion. Groupings of three nucleotides indicate the reading frame of lacZa. The Taq Pol I gene would be minimal. Attempts to clone the gene under control of its own promoter in E. coli were unsuccessful, probably because of the detrimental effect the polymerase had on the cell. We did not know if Taq Pol I would be toxic to E. coli cells at 37 °C. While the in vitro specific activity of Taq Pol I at 37 °C is only a few percent of the specific activity at 75-80 °C, we could not predict if the DNA binding activity of the enzyme might interfere with normal cell function. To avoid potential problems related to direct expression of the gene in E. coli we chose to clone an epitope of the Taq Pol I gene by using λgt11 libraries and antibody selection. The epitope-expressing clone was subsequently used to select the entire Taq Pol I gene from a library in λCh35.

We were unable to detect a thermostable polymerase activity in cells infected (11) with any of the λCh35 clones, including p42. The polymerase assay is extremely sensitive and can detect 1 molecule of polymerase per 10 cell equivalents. Upon subcloning of the 8-kb probe-hybridizing HindIII A fragment from p42 into BSM13 and IPTG induction of the subclone pFC82.35, a low level of thermostable polymerase activity was detected (Table I). Based on the activity of purified Taq Pol I when admixed with E. coli cells, this activity represents two to three molecules of Taq Pol I per cell equivalent. The gene was localized to one end of the 8.0-kb HindIII A fragment by using deletion analysis. Upon IPTG induction, pFC84, the SphI deletion, and pFC85, the AseI deletion, yielded a 100-fold increase in Taq Pol I activity (Table I) compared to that of the full-length A fragment subclone, pFC82.35. This increase in activity allowed for ready detection of the induced protein(s) on Western blot (Fig. 5). The A fragment induced proteins were truncated with an

**TABLE II**

| Fusion* | LacZa phenotypeb | Fusion DNA sequencec |
|---------|------------------|----------------------|
| ΔNe 1   | Blue             | GAG CTA G             |
| ΔBa 15  | White            | CAGAG047              |
| ΔBa 53  | Blue             | GGG CAG AGG ATC       |
| ΔBa 35  | Blue             | GGG CAG AGG ATC       |
| Δxho 2b | White            | TTCGCCGGCTCTG         |
| Δxho 30 | White            | CGTGCGGGAGGGATCG      |
| Δxho 53 | Blue             | GAA GGC CTT GGC       |
| Δxho 54 | Blue             | GAG GGG GTG GCC       |
| Δxho 59 | Blue             | GAG GGG GGG GGG       |

* Construction of fusions between 5' sequences of the Taq Pol I A fragment and lacZa is described under “Materials and Methods.”

b The lacZa phenotype was determined on agar plates containing X-Gal. In-frame fusions resulted in blue colonies on X-Gal and out-of-frame fusions yielded white colonies.

c The DNA sequence was determined at the site of each fusion. Groupings of three nucleotides indicate the reading frame of lacZa. The Taq Pol I gene would be minimal. Attempts to clone the gene under control of its own promoter in E. coli were unsuccessful, probably because of the detrimental effect the polymerase had on the cell. We did not know if Taq Pol I would be toxic to E. coli cells at 37 °C. While the in vitro specific activity of Taq Pol I at 37 °C is only a few percent of the specific activity at 75-80 °C, we could not predict if the DNA binding activity of the enzyme might interfere with normal cell function. To avoid potential problems related to direct expression of the gene in E. coli we chose to clone an epitope of the Taq Pol I gene by using λgt11 libraries and antibody selection. The epitope-expressing clone was subsequently used to select the entire Taq Pol I gene from a library in λCh35.

We were unable to detect a thermostable polymerase activity in cells infected (11) with any of the λCh35 clones, including p42. The polymerase assay is extremely sensitive and can detect 1 molecule of polymerase per 10 cell equivalents. Upon subcloning of the 8-kb probe-hybridizing HindIII A fragment from p42 into BSM13 and IPTG induction of the subclone pFC82.35, a low level of thermostable polymerase activity was detected (Table I). Based on the activity of purified Taq Pol I when admixed with E. coli cells, this activity represents two to three molecules of Taq Pol I per cell equivalent. The gene was localized to one end of the 8.0-kb HindIII A fragment by using deletion analysis. Upon IPTG induction, pFC84, the SphI deletion, and pFC85, the AseI deletion, yielded a 100-fold increase in Taq Pol I activity (Table I) compared to that of the full-length A fragment subclone, pFC82.35. This increase in activity allowed for ready detection of the induced protein(s) on Western blot (Fig. 5). The A fragment induced proteins were truncated with an

**FIG. 6. Plasmid pLSG1.** The 6.58-kb plasmid contains a 3.41-kb HindIII A fragment and lacZa is described under “Materials and Methods.” The DNA sequence was determined at the site of each fusion. Groupings of three nucleotides indicate the reading frame of lacZa. The Taq Pol I gene would be minimal. Attempts to clone the gene under control of its own promoter in E. coli were unsuccessful, probably because of the detrimental effect the polymerase had on the cell. We did not know if Taq Pol I would be toxic to E. coli cells at 37 °C. While the in vitro specific activity of Taq Pol I at 37 °C is only a few percent of the specific activity at 75-80 °C, we could not predict if the DNA binding activity of the enzyme might interfere with normal cell function. To avoid potential problems related to direct expression of the gene in E. coli we chose to clone an epitope of the Taq Pol I gene by using λgt11 libraries and antibody selection. The epitope-expressing clone was subsequently used to select the entire Taq Pol I gene from a library in λCh35.

We were unable to detect a thermostable polymerase activity in cells infected (11) with any of the λCh35 clones, including p42. The polymerase assay is extremely sensitive and can detect 1 molecule of polymerase per 10 cell equivalents. Upon subcloning of the 8-kb probe-hybridizing HindIII A fragment from p42 into BSM13 and IPTG induction of the subclone pFC82.35, a low level of thermostable polymerase activity was detected (Table I). Based on the activity of purified Taq Pol I when admixed with E. coli cells, this activity represents two to three molecules of Taq Pol I per cell equivalent. The gene was localized to one end of the 8.0-kb HindIII A fragment by using deletion analysis. Upon IPTG induction, pFC84, the SphI deletion, and pFC85, the AseI deletion, yielded a 100-fold increase in Taq Pol I activity (Table I) compared to that of the full-length A fragment subclone, pFC82.35. This increase in activity allowed for ready detection of the induced protein(s) on Western blot (Fig. 5). The A fragment induced proteins were truncated with an
apparent molecular mass of 63–65 kDa. Fusing the 5' HindIII site in the A fragment with the HindIII site in BSM13* causes the Taq Pol I gene to be out of frame with respect to β-galactosidase. The reading phase at the HindIII site in BSM13* with respect to β-galactosidase is A AGC TT, a frame of "0" (13). The reading frame of Taq Pol I at the HindIII site is AAG CTT ("plus 1"). The fusion gives rise to a minus 1 frame shift. In the β-galactosidase reading frame, there is a TGA stop codon at nucleotide 1478 of Taq Pol I. Downstream of this TGA there are several possibilities for restarts which could result in truncated forms of Taq Pol I: ATGs at nucleotides 1509 and 1752 and GTGs at nucleotides 1547, 1569, 1722, and 1731. In fact, we see a doublet in induced lanes of both pFC84 and pFC85 on Western blots (Fig. 5) indicating at least two reinitiation sites. All but one of the likely sites, the ATG at nucleotide 1809, would probably require a ribosome binding site for reinitiation. There are reasonable ribosome binding sites for the GTG at nucleotide 1722 and for the ATG at nucleotide 1752. Translation initiating at these sites would yield proteins of 59 and 58 kDa, respectively. However, the apparent molecular masses of the doublet bands seen on Western blots of pFC84 and pFC85 are approximately 65 and 63 kDa, based on comparison of the mobilities of the doublet bands with the molecular weight size standards. Whether the result of reinitiation or proteolytic processing, the thermostable, enzymatically active, truncated forms of Taq Pol I directed by plasmids pFC84 and pFC85 (Table I) suggest that significant portions of the Taq Pol I sequence are not essential for DNA polymerase activity.

The purpose of the set of fusions of 5' portions of the Taq Pol I A fragment with lacZα at BSM13* was to confirm or determine the reading phase of the Taq Pol I gene internally as an aid to nucleotide sequencing. Since we knew the reading phase of lacZα at BSM13* polylinker, we could infer the reading phase of Taq Pol I in α-complementing in-frame fusions. DG98 harboring fusions which were in-frame were readily detectable as blue colonies on X-Gal indicator plates. We generated a series of fusions (Table II) at nine sites between nucleotides 962 and 1782 of the Taq Pol I gene.

We compared the DNA sequence of Taq Pol I with that of E. coli DNA polymerase I. At the DNA level, the two genes lack any significant regions of homology (Table III). In regions where the amino acid sequences are homologous, the DNA sequences diverge, especially in third positions of codons. The longest stretch of DNA sequence identity is 19 bases (Table III).

The predicted amino acid sequence of Taq Pol I is shown in Fig. 2. From this a codon bias table was generated (Table IV). There is a heavy bias toward G and C in the third position (91.8% C and G) as would be expected for GC-rich organisms and as others have observed for other Thermus genes: 95% C and G and for the gk24 gene encoding L-lactate dehydrogenase of Thermus caldophilus (15), 94.8% for mdh from T. flavus (14), and 89% for leuB from T. thermophilus (16).

Significant amino acid sequence similarity exists between Taq Pol I, E. coli Pol I, and bacteriophage T7 DNA polymerase. One possible sequence alignment yields 38% identity between the Taq Pol I and E. coli Pol I amino acid sequences (Fig. 7). There are two major regions of Taq Pol I and one region of T7 DNA polymerase that show extensive sequence similarity compared to E. coli Pol I. The first region of Taq Pol I extends from the N terminus to approximately residue 300. The second region extends from approximately residue 410 to the C terminus of Taq Pol I. The N-terminal region of Taq Pol I corresponds to the N-terminal domain of E. coli Pol I shown to contain the 5'-3' exonuclease activity (17). The C-terminal regions of Taq Pol I and T7 DNA polymerase correspond to the E. coli Pol I domain shown to contain DNA polymerase activity (18). The x-ray structure of the Klenow fragment (19) shows that this domain contains a deep cleft believed to be responsible for DNA binding.

Accordingly as a result of many mutations, deletions, insertions, etc. during evolution, Taq Pol I residues at positions 300–410 show little sequence similarity compared to E. coli Pol I. Taq Pol I is 96 residues shorter than E. coli Pol I; most of the deleted residues occur in the region encompassing residues 300–410. Ollis et al. (19) and Derbyshire et al. (20) have shown that E. coli Pol I residues Asp-355, Glu-357, Leu-361, Asp-424, Phe-473, and Asp-501 are involved in binding of divalent cation and deoxynucleoside monophosphate. A fragment of E. coli Pol I that contains only residues 515–928 is devoid of 3'-5' exonuclease activity, but still retains polymerase activity (18). Presumably, the E. coli Pol I region comprised of residues 324–515 forms at least part, if not all, of the 3'-5' exonuclease activity. Taq Pol I and E. coli Pol I display little sequence similarity in the presumptive 3'-5' exonuclease region. Of the E. coli Pol I residues shown to be involved in cation and deoxynucleoside monophosphate binding, the sequence alignment of Fig. 7 shows only Asp-424 as having an exact homolog in the Taq Pol I sequence. Although other high scoring sequence alignments are possible in the Taq Pol I 300–410 region, it is possible that the Taq Pol I gene has undergone key mutations, deletions, or insertions

## TABLE III

| DNA sequence identity of Taq Pol I and E. coli Pol I |
|-----------------------------------------------|
| Sequence location* | Nucleotide identity | Amino acid identity |
| Taq Pol I | 190–205 | 178–196 | 19/19 | 6/6 |
| Taq Pol I | 1730–1757 | 2015–2042 | 23/28 | 9/9 |
| Taq Pol I | 2200–2277 | 2545–2562 | 17/18 | 6/6 |
| Taq Pol I | 2344–2363 | 2635–2654 | 17/20 | 7/7 |

* Nucleotide sequence coordinates for Taq Pol I from Fig. 2. Nucleotide sequence coordinates for E. coli Pol I adapted from GenBank.

## TABLE IV

| Codon usage in the T. aquaticus DNA polymerase I gene |
|-----------------------------------------------|
| Codon | Usage | Percentage |
| Arg | CTT | 0/127 | Ser |
| CCG | 24 (124) | 30 (31) | TCT |
| CGG | 27 | CTT | 20 | TGC |
| CGA | 0 | CTC | 46 | TCA |
| AGG | 25 | CTG | 50 | AGT |
| AGA | 0 | CTA | 5 | AGC |
| Thr | ACT | 0 | Pro | CTT | 3 |
| ACC | 10 | CCC | 34 (51) | GTG |
| ACA | 0 | CCA | 2 | GTA |
| Ala | GCT | 2 | Gly | GGT | 0 | Ile |
| GCC | 77 (58) | GGG | 28 (25) | ATC |
| GCA | 12 | GGC | 30 | ATA |
| GCA | 0 | GGA | 0 |
| Aan | AAT | 0 | Gln | CAG | 15 | Tyr |
| ACC | 12 (16) | CAA | 1 (24) | TAC |
| His | CAT | 0 | Gln | GAG | 79 | Cys |
| ACC | 18 (87) | GAA | 8 | TGC |
| Asp | GAT | 3 | Phe | TT  | 8 | Lys |
| GAC | 39 (27) | TTC | 19 (42) | AAA |
| Met | ATG | 16 | Trp | TGG | 14 |

| Codon | Usage | Percentage |
|-----------------------------------------------|
| Arg | CTT | 0/127 | Ser |
| CCG | 24 (124) | 30 (31) | TCT |
| CGG | 27 | CTT | 20 | TGC |
| CGA | 0 | CTC | 46 | TCA |
| AGG | 25 | CTG | 50 | AGT |
| AGA | 0 | CTA | 5 | AGC |
| Thr | ACT | 0 | Pro | CTT | 3 |
| ACC | 10 | CCC | 34 (51) | GTG |
| ACA | 0 | CCA | 2 | GTA |
| Ala | GCT | 2 | Gly | GGT | 0 | Ile |
| GCC | 77 (58) | GGG | 28 (25) | ATC |
| GCA | 12 | GGC | 30 | ATA |
| GCA | 0 | GGA | 0 |
| Aan | AAT | 0 | Gln | CAG | 15 | Tyr |
| ACC | 12 (16) | CAA | 1 (24) | TAC |
| His | CAT | 0 | Gln | GAG | 79 | Cys |
| ACC | 18 (87) | GAA | 8 | TGC |
| Asp | GAT | 3 | Phe | TT  | 8 | Lys |
| GAC | 39 (27) | TTC | 19 (42) | AAA |
| Met | ATG | 16 | Trp | TGG | 14 |
that have destroyed its 3'-5' exonuclease activity. Preliminary results indicate that Taq Pol I displays little if any 3'-5' exonuclease activity.

Sequence homology between \(E. coli\) Pol I and \(T. aquaticus\) has been previously noted. Those \(T.7\) DNAs that have destroyed its exonuclease activity. Preliminary data scoring matrix of Staden and Pizzagalli (22, 23). A complete and unambiguous sequence alignment of \(T. aquaticus\) and \(E. coli\) Pol I are also present in the Taq Pol I and \(T.7\) DNA polymerase.

Sequence homology between \(E. coli\) Pol I and \(T.7\) DNA polymerase has been previously noted. Those \(T.7\) DNA polymerase sequences shown by Ollis et al. (21) to be conserved between that enzyme and \(E. coli\) Pol I are also present in the Taq Pol I amino acid sequence (Fig. 7). Most of the conserved residues are found in structural features that form the DNA-binding cleft of the enzyme. Although short segments of \(T.7\) DNA polymerase sequence in the 1–334 region are similar to regions in \(E. coli\) Pol I and Taq Pol I, the overall sequence similarity in this region, ignoring the first 300 residues of \(E. coli\) Pol I and Taq Pol I that form the 5'-3' exonuclease domain, is poor. A complete and unambiguous sequence alignment for this region cannot be assigned. It should be noted that although \(T.7\) DNA polymerase also shows little similarity to \(E. coli\) Pol I in the region of the 3'–5' exonuclease domain, \(T.7\) DNA polymerase is reported to display significant 3'–5' exonuclease activity (22, 23).

Bernard et al. (24) and Pizzagalli et al. (25) have identified several short regions of DNA polymerase amino acid se-
quences that are highly conserved. The conserved sequences are found in polymerases from herpes simplex virus type 2, human cytomegalovirus, varicella-zoster virus, Epstein-Barr virus, vaccinia virus, adenovirus type 2, killer plasmid from Kluyveromyces lactis, maize mitochondrial particle, bacterio-

that hold true, it is particularly notable that the Arg to Lys or Ala or Ala amino acid preferences in proteins from thermophilic orga-

Chemical modification and inactivation studies of *E. coli* Pol I have resulted in the identification of many amino acid residues believed to be important or essential for polymerase activity (26-31). Among these residues are: Met-512, Arg-682, Lys-758, Tyr-766, Arg-841, and His-881. Comparing the Taq Pol I amino acid sequence to the *E. coli* Pol I sequence, all of the above residues, except Met-512, are conserved. Taq Pol I contains a Leu residue at the analogous position. Apparently, the functionally similar Taq Pol I Leu residue at position 417 can fulfill the role ascribed to *E. coli* Pol I Met-512 in template primer binding (30).

Analyses on the effects of various mutations in the *E. coli* Pol I gene upon enzymatic activity have also been used to define amino acid residues important for polymerase activity. For example, a Gly to Arg mutation at position 850 (*polA5*) results in a polymerase that is less processive on the DNA substrate (32). An Arg to His mutation at position 690 (*polA6*) results in a polymerase that is defective in DNA binding (33). Both Gly-850 and Arg-690 are conserved residues in Taq Pol I. Joyce et al. (34) have characterized a number of *E. coli* Pol I mutants defective in 5'-3' exonuclease activity. Interestingly, the four mutations, Y77C (*polA107*), G184D (*polA1113*), G184D (*polA480ex*), and G192D (*polA214*) all occur at amino acid residues that are conserved in Taq Pol I.

As would be expected for an enzyme from a thermophilic organism, Taq Pol I is considerably more thermostable than Pol I from *E. coli* (data to be presented in a later publication). Although a better assessment of an enzyme's thermostability would result from a complete cataloging of all stabilizing amino acid interactions, in the absence of high resolution x-ray crystal structures, many researchers have attempted to explain enzyme thermostability by an analysis of amino acid content (35-37). Several features of thermostable enzymes have been noted in such studies. Among those features are increased ratios of Arg to Lys residues, Gln to Asp residues, Ala to Gly residues, Thr to Ser residues, and a reduced Cys content. Comparing Taq Pol I to *E. coli* Pol I, the Ala to Gly and Thr to Ser ratios are smaller for Taq Pol I than for *E. coli* Pol I. Of the thermostabilizing type amino acid alterations that hold true, it is particularly notable that the Arg to Lys ratio for Taq Pol I is nearly twice that for *E. coli* Pol I. It is possible that the propensity of thermophilic proteins to contain Arg rather than Lys residues is simply a reflection of the high GC content of thermophilic organisms. The structural gene for Taq Pol I contains 67.9% GC compared to a 52.0% GC content for *E. coli* Pol I. The six Arg codons are rich in C and C (13 out of 18 bases are G or C) compared to the two Lys codons (1 out of 6 bases is G). This explanation for amino acid preferences in proteins from thermophilic organisms cannot be the basis for Gln versus Asp, Thr versus Ser, or Ala versus Gly preference, because there are equal ratios of GC versus AT in the codons for those pairs of amino acids.

A more likely explanation for the preference for Arg over Lys in thermostable proteins would seem to be based on the unique physical-chemical properties of the two amino acids (e.g. pK values, hydrogen bonding patterns, hydrophobicity/hydro-philicity).

The truncated and full-length Taq Pol I enzymes produced upon IPTG induction show different reactivities to the anti-

The level of expression in *E. coli* of full-length Taq Pol I encoded by pLSG1 is similar to the level of expression of Taq DNA polymerase in *T. aquaticus*. In pLSG1 (Fig. 6) the beginning of the Taq Pol I open reading frame is 109 bp distal to the BglII site and 171 bp distal to the lacZα translation initiation site. A low level of Taq Pol I expression in cells harboring pLSG1 is consistent with an in-phase TGA codon (~111 through ~109, Fig. 2) in the Taq DNA sequence causing translation termination of the lacZα polypeptide. Reinitiation of translation at the first ATG results in the synthesis of the 94-kDa Taq Pol I protein. Further manipulation of the Taq DNA polymerase sequence has increased the level of expression. The cloned full-length Taq Pol I gene in pLSG1 affords the advantages of expressing Taq Pol I in *E. coli* and in ease of isolating the enzyme from *E. coli* compared to *T. aquaticus*. These advantages will aid in further study of the enzyme and will provide a ready source of Taq Pol I for use in PCR and other biochemical procedures in which Taq Pol I might prove useful, such as in DNA sequencing.

Acknowledgments—We gratefully acknowledge Gail Rodgers for advice on preparation of antibody from a small amount of protein; Corey Levenson, Lauri Goda, and Dragan Spasic for preparation of oligonucleotide primers, Keith Bauer for other biochemical procedures in which Taq Pol I might prove useful, such as in DNA sequencing.

REFERENCES

1. Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985) *Science* 230, 1350-1354

2. Mullis, K. B., and Faloona, F. A. (1987) *Methods Enzymol.* 155, 335-350

3. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988) *Science* 239, 487-491

4. Chien, A., Edgar, D. B., and Trela, J. M. (1986) *Bacteriol. Rev.* 127, 1550-1557

5. Kaledin, A. G., Snyazovskaya, A. G., and Gorodetskii, D. P., and Matvienko, V., Vorozheykina, D. P., and Matvienko, V., Vorozheykina, D. P., and Matvienko, V. (1988) *Nucleic Acids Res.* 16, 1563-1575

6. Murzina, N. V., Vorozheikina, D. F., and Matvienko, N. I. (1988) *Nucleic Acids Res.* 16: 8172

* F. C. Lawyer, Stoffel, and D. H. Gelfand, unpublished observations.
Isolation and Expression of Taq Pol I Gene

SUPPLEMENTARY MATERIAL TO

Identification, Characterization and Expression in E. coli of the DNA Polymerase Gene from Thermoanaerobacterium thermosulfurigenes

Frances C. Lawvere, Susan L. Wiltz, Ronald R. Sank, Kesten L. Weinbach, and John W. Giddings

MATERIALS AND METHODS

Restriction Enzymes - The restriction enzyme was Y17, AEC 176106, which was obtained from The American Type Culture Collection (ATCC). Restriction endonuclease activity was determined by the method described by Giddings et al. (10).

Isolation and Expression of Taq Pol I Gene - Restriction enzyme digestion was performed as follows: 1.5 units of restriction enzyme was added to a volume of 100 units of Taq Pol I DNA in 20 mM Tris-acetate (pH 7.0), 10 mM MgCl₂, and 50 mM NaCl. The mixture was incubated at 37°C for 45 minutes. The enzyme was then inactivated by heating at 65°C for 5 minutes. The digested DNA was purified by phenol extraction and ethanol precipitation.

RESULTS

Cloning Restriction Fragments - Daughter restriction enzymes were purified from New England Biolabs, Inc., and NheI was added to the digested Taq Pol I DNA. The resulting digest was then purified by gel electrophoresis and then ligated into the plasmid vector pBR322. The ligated DNA was then transformed into E. coli strain HB101. The transformed cells were then grown on Luria Bertani (LB) agar plates containing 25 mM ampicillin and 40 mM kanamycin.

Expression of Taq Pol I Gene - The plasmid pUC19-Taq Pol I was then isolated from the transformed E. coli strain HB101 and purified by alkaline lysis. The purified plasmid DNA was then transformed into E. coli strain S17-1, which is a derivative of the lacZ strain E. coli strain HB101. The transformed cells were then grown on LB agar plates containing 25 mM ampicillin and 40 mM kanamycin. The resulting colonies were then assayed for beta-galactosidase activity using the X-gal assay.

DISCUSSION

The results presented in this paper demonstrate the feasibility of isolating and expressing the Taq Pol I gene in E. coli. The Taq Pol I gene was isolated from the genome of Thermoanaerobacterium thermosulfurigenes and was found to encode a protein with significant homology to the Taq Pol I enzyme. The expression of the Taq Pol I gene in E. coli resulted in the production of a protein that was similar to the Taq Pol I enzyme. These results suggest that the Taq Pol I gene can be used as a model system for the study of DNA polymerase function and regulation.
### Table V
DNA Sequencing Primers

| Primer | Taq Pol I Nucleotide Coordinates |
|--------|----------------------------------|
| MK122  | 828 → 807                        |
| MK123  | 820 → 808                        |
| MK128  | 861 → 879                        |
| MK130  | 2108 → 2127                      |
| MK131  | 1873 → 1897                      |
| MK132  | 1825 → 1841                      |
| MK133  | 3223 → 3240                      |
| MK134  | 3388 → 3406                      |
| MK135  | 293 → 300                        |
| MK136  | 295 → 302                        |
| MK138  | 345 → 352                        |
| MK139  | 1013 → 1022                      |
| MK140  | 125 → 136                        |
| MK141  | 1071 → 1084                      |
| MK143  | 1115 → 1122                      |
| MK144  | 1355 → 1362                      |
| MK145  | 2360 → 2379                      |
| MK146  | 2410 → 2429                      |
| MK150  | 318 → 327                        |
| MK155  | 2501 → 2520                      |
| MK156  | 297 → 308                        |
| MK159  | 48 → 48                          |

*Arrows drawn sequenced with each primer. → is sense strand, 5' to 3', and ← is antisense strand, 3' to 5'.

---

**Fig. 6. DNA sequencing strategy.** The bold line on the map delineates the coding sequence for Taq Pol I. Arrows indicate sequence obtained on the sense (→) or antisense (←) strand. Length of the arrows corresponds to the amount of sequence obtained in each case. 100% of the DNA sequence was determined on each strand. 1/3, 2/3, the remaining sequencing primer was used for sequencing on templates pC212, pC203, and the B fragments bigill bigill. 4/1, the reverse sequencing primer was used with a fragment isolation Xh Xh . The 5' to 3' reverse sequencing primer was used with A fragment isolation Xh Xh Sh. Primer annealed were MK122-126, MK130-136, MK134-145, MK143-148, MK153-154, MK157, MK158, and MK159.