A DNA polymerase from the archaeon *Sulfolobus solfataricus* shows sequence similarity to family B DNA polymerases

Francesca M. Pisani, Ciro De Martino and Mosè Rossi*
Istituto di Biochimica delle Proteine ed Enzimologia, Consiglio Nazionale Ricerche, Via Marconi 10, 80125, Naples, Italy

Received March 18, 1992; Revised and Accepted May 7, 1992
EMBL accession no. X64466

ABSTRACT
The gene encoding the thermostable DNA polymerase from the archaeon *Sulfolobus solfataricus* (strain MT 4) was isolated by means of two degenerate oligonucleotide probes. They were designed on the basis of partial enzyme amino acid sequences. The gene was found to encode a 882 residues polypeptide chain with a deduced molecular mass of about 100 kDa. By comparison with other archaeal genes, putative regulatory sites were identified in the gene-flanking regions. By computer-assisted homology search, several sequence similarities among *S. solfataricus* and family B DNA polymerases were found. In addition, conserved sequence motifs, implicated in the 3'-5' exonuclease activity of *E. coli* DNA polymerase I and shared by various family A and B DNA polymerases, were also identified. This result suggests that the proofreading domains of all these enzymes are evolutionarily related.

INTRODUCTION
The study on evolution and structural relatedness of DNA polymerases is of particular interest because of their essential and almost universal role in the transmission of genetic information from one generation to the next (1). Based on primary structural similarities, DNA polymerases have been recently classified into at least two major families (e. g. 2). The family A includes DNA polymerase I (pol I) from *E. coli* (3), DNA polymerases from *S. pneumoniae* (4), *T. aquaticus* (5), and bacteriophages T5 (6), T7 (7) and SPO2 (8). They are all very sensitive to dideoxynucleotides and resistant to aphidicolin. On the other hand, the family B is quite heterogeneous, including human (9) and yeast (10) DNA polymerases α, *E. coli* DNA polymerase II (11), DNA polymerases from eukaryotic viruses (12–14) and from bacteriophage T4 (15) and protein-primed replicases (16). Most of the family B enzymes are sensitive to aphidicolin and relatively resistant to inhibition by dideoxynucleotides.

Obviously, the final aim of these classifications, apart from evolutionary implications, is the identification of conserved sequence segments that are presumably essential to the biological functions of these enzymes. Sequence data on archaeal DNA polymerases could provide sufficient variability in order to recognize the functionally critical regions. Indeed, Archaea (recently renamed Archaea) define a primary domain, which is at molecular level as distant from that of Eubacteria (or Bacteria) as both are from that of Eukaryotes (or Eucarya) (17). In this context, we undertook the cloning of DNA polymerase gene from the thermoacidophilic archaeon *Sulfolobus solfataricus* (18). This enzyme was shown to be a monomer of about 100 kDa, very sensitive to proteinase, highly thermophilic and related to α-like DNA polymerases because of its sensitivity to aphidicolin and resistance to inhibition by dideoxynucleotides (19,20).

We present here the sequence of *S. solfataricus* DNA polymerase gene (pol S) and the deduced primary structure of its protein product. To our knowledge, this is the first report on a complete gene sequence of an archaeal DNA polymerase. Several regions conserved among *S. solfataricus* and class B DNA polymerases were identified; their putative functional roles are discussed.

EXPERIMENTAL PROCEDURES
Enzymes, radioactive biochemicals and synthetic oligonucleotides
Restriction and modification enzymes were purchased from Promega. They were used according to standard procedures or as recommended by the supplier. Radioactive biochemicals were obtained from Amersham. The oligonucleotides used in this study were synthesized by conventional solid phase methods and purified by gel filtration.

Bacterial strains, genomic libraries and cloning vectors
Cells of *Sulfolobus solfataricus* (strain MT 4) were grown as previously described (18). *Escherichia coli* strain Y1090 (21) and C600 hfl (22) were used for plating Agt11 and λEMBL3A phages
respectively. *E. coli* strain DH5α was used for the propagation of plasmids (22). λgt11 and pEMBL8A genomic DNA libraries were previously described (23). Plasmids pUC18 (24) and pEMBL8 (25) were used for subcloning and DNA sequencing by standard techniques (22).

**Amino acid microsequence analysis of DNA polymerase**

A sample of highly purified DNA polymerase (30 mg) was loaded onto a 10% SDS-polyacrylamide gel. After the electrophoretic run, the gel was electroblotted onto Polyvinylidene difluoride (PVDF) membrane ProBlott (Applied Biosystems), as described (26). The membrane was stained with Coomassie Brilliant Blue R-250 and destained following manufacturer's instructions. As previously reported (19,20), three major protein bands with molecular masses of about 100, 50 and 40 kDa were detected. Each of them was cut out by a sterile razor blade and loaded directly onto an Applied Biosystems gas-phase sequencer (Model 470 A) equipped with an on-line PTH-analyzer. The N-terminal residue of 100 kDa protein was found to be blocked directly onto an Applied Biosystems gas-phase sequencer (Model 470 A) equipped with an on-line PTH-analyzer. The N-terminal residue of 100 kDa protein was found to be blocked to Edman degradation, while we succeeded in determining the terminal residue of 100 kDa protein was found to be blocked to Edman degradation, while we succeeded in determining the amino-terminal sequence of 40 and 50 kDa polypeptides, which were (in the single letter amino acid code): SAPVEEEKVVR and GYKGAVVVIDP, respectively.

**Isolation and sequencing of *S. solfataricus* DNA polymerase gene**

According to *S. solfataricus* codon usage (23,27), two degenerate sense oligonucleotides were designed: the 23-mer SSDP40k 5'-GC(T/A) CC(T/A) GT(T/A) GA(A/G) GA(A/G) AA(A/G) GT-3', corresponding to the amino-terminal sequence of the 40 kDa polypeptide; the 29-mer SSDP50k 5'-GGA TA(T/C) GG(T/A) GG(T/A) GC(T/A) GT(T/A) AT(T/A) GAT CC-3', corresponding to the amino-terminal sequence of the 50 kDa polypeptide. Each oligonucleotide, 5'-end labeled with T4 polynucleotide kinase and [γ-32P]ATP, was used separately to screen a *S. solfataricus* genomic library in λgt11. Filter hybridizations were carried out at 45°C with the SSDP40k probe and at 50°C with the SSDP50k probe in 6x sodium saline citrate buffer, according to Maniatis (22). The EcoRI-inserts of the positive phages, subcloned into pUC18 vector, were characterized by restriction mapping and cross-hybridization analysis. The subclones pE1 (hybridizing to probe SSDP40k), pA1 and pC5 (hybridizing to probe SSDP50k) were also sequenced by dideoxy chain termination method using T7 DNA polymerase (Sequenase,USB). They were found to overlap (see Fig.1) and to contain an open reading frame (ORF) encoding the protein sequence utilized to design the respective oligonucleotide probes. In order to isolate the full-lenght DNA polymerase gene, we screened with the Eco RI-insert of subclone pC5 as a probe the Sulfolobus genomic library in λEMBL3. This library contains inserts of about 15 kb obtained by Mbo I partial digestion of *S. solfataricus* genomic DNA. The filter hybridizations were carried out at 65°C in 6x sodium saline citrate buffer, according to Maniatis (22). Two positive phages (λA4B and λA2P) were purified and analyzed by restriction mapping (Fig. 1). A common 5kb EcoRI-PstI fragment, hybridizing to the λgt11 subclones pE1, pA1 and pC5, was inserted into pEMBL8 vector to generate the plasmid pFCpolS. Sequencing reactions were performed on this plasmid mostly utilizing specific synthetic primers (Fig.1). All sequences were determined in duplicate on both strands and analyzed using MicroGenie DNA Computer program (Beckman).

**Computer analysis**

Protein similarity searches were performed on a Vax computer using FASTA program (28) on Swissprot (Release n.18) and Protein Identification Resource (Release n.29) databases. Multiple alignments were obtained using PILEUP program (29) and manually corrected to achieve maximum homology.

**RESULTS AND DISCUSSION**

**Isolation and nucleotide sequence of *S. solfataricus* DNA polymerase gene**

DNA polymerase from *S. solfataricus* was previously shown to have a native molecular mass of about 100 kDa by gel filtration chromatography and glycerol gradient centrifugation. Denaturing polyacrylamide gel electrophoresis of the homogeneous enzyme gave, in addition to the 100 kDa protein band, two major polypeptides of about 50 and 40 kDa. They were thought to be proteolytic fragments of the 100 kDa protein, because Western blot analysis with a polyclonal antiserum against the intact 100 kDa polypeptide showed their immunological relatedness. Furthermore, by activity gel analysis according to Karawya (30), in addition to the 100 kDa catalytic species, also the 50 kDa polypeptide was found to retain DNA polymerase activity. All that considered, we performed amino acid sequence analysis of the above polypeptides after their separation through a SDS-polyacrylamide gel and electrotransfer to a PVDF membrane. Attempts to determine the 100 kDa polypeptide sequence were

---

**Fig. 1.** Physical map of Pol S gene and its flanking regions. The lower part of the figure represents the overlap of subclones from λgt11 positive phages isolated with SSDP40k or SSDP50k oligonucleotide probe. The upper part shows the restriction map of the insert from the positive λEMBL3 phage 4B and of subclone pFCpolS. The DNA sequencing strategy is schematically described: arrows with circles and without circles represent regions sequenced with a universal primer on subclones from λgt11 positive phages and regions sequenced with synthetic primers on plasmid pFCpolS, respectively.
unsuccessful, suggesting that its amino-terminal residue is blocked with an acyl group, such as acetyl. On the other hand, we succeeded in determining the amino-terminal sequence of the 40 and 50 kDa fragments. Two degenerate oligonucleotides, based upon this sequence information, were utilized to isolate the DNA polymerase gene from *S. solfataricus* genomic libraries in phage vectors.

Fig. 2 shows the nucleotide sequence of this gene determined from the subclone pFCpolS by the strategy schematically described in Fig. 1. A 882-codon ORF starting with ATG at position 198 and ending with a TAA stop signal at position 2844 is found within this sequence. It corresponds to a protein with a predicted molecular mass of about 100 kDa, in good agreement with that reported for *S. solfataricus* DNA polymerase. The amino-terminal sequences of the 50 and 40 kDa polypeptides are entirely encoded within this ORF (Fig. 2). These findings confirm the identity of the gene and demonstrate the occurrence of a proteolytic degradation of the enzyme during the purification procedure. Since the amino-terminal residue of the 40 kDa polypeptide is a Serine, a proteolytic cleavage of the native protein at this site must also be assumed. The ATG at position 198 is the unique possible initiation codon located upstream of the triplet for the amino-terminal Serine, before a stop signal is encountered. Indeed, the comparison of the region proximal to this putative translation starting point with the 3'-end of 16S RNA from *S. solfataricus* (31) revealed the presence of a potential ribosome

![Fig. 2. Nucleotide sequence of Pol S gene and deduced protein primary structure. The underlined amino acid sequences were utilized to design the probes for Agt11 library screening. Putative DNA regulatory elements are underlined.](image-url)
binding site overlapping the above ATG codon, as reported for some other archaeal genes (23, 28, 31). The 5' non-coding region of the PolS gene does not seem to possess a sequence compatible with the consensus drawn for the archaeal promoters (32–33). This could suggest that it is part of a multigene transcriptional unit. On the other hand, downstream from the termination codon a pyrimidine-rich region containing the heptanucleotide sequence TTTTTAT was identified, perfectly matching to the structure of archaeal terminators (23, 27, 34).

Sequence similarities with other DNA polymerases and assessment of functional domains

By computer-assisted comparison of S. solfataricus DNA polymerase primary structure with Swissprot and Protein Identification Resource data banks we found several similarities with family B DNA polymerases, including eukaryotic-viral replicates (12–14), human (9) and S. cerevisiae (10) DNA polymerases α. On the other hand, few similarities were detected with E. coli pol I (3).

In class B DNA polymerases conserved sequence motifs have been found having the same relative location within the respective polypeptide chains (2, 9, 11, 35–37). These conserved regions were also identified in the amino acid sequence of S. solfataricus DNA polymerase in the same spatial arrangement (Regions 1–8 in Fig. 3).

The Regions 1, 2 and 3 correspond to the EXO sequence motifs shared by a variety of DNA polymerases with 3'-5' exonuclease activity. They were recently aligned by Morrison et al. (38) to E. coli pol I amino-terminal regions containing the critical residues Asp 355, 424 and Glu 357. These aspartic acid residues were shown in the Klenow fragment crystals to anchor two divalent metal ions that are essential for exonuclease activity, whereas the Glutamic acid 357 is supposed to play a separate role in catalysis of this reaction (39–40). Furthermore, site-directed mutagenesis experiments of yeast (38) and Φ29 (41) DNA polymerase residues corresponding to the E. coli Asp 355 and Glu 357 were found to result in the loss of exonuclease activity and/or in a mutator phenotype, consistent with a function in proofreading. The multiple sequence alignment of Fig. 3 seems to indicate that all the above carboxylate residues are also conserved in the amino-terminal portion of the archaeal DNA polymerase. This result concurs with our finding that Sulfobolus DNA polymerase has an associated 3'-5' exonuclease activity, as recently determined by 3'-terminal mismatch excision assay (M. P., unpublished result).

Although the sequence similarities Regions 4 to 8 of Fig. 3 are highly conserved in family B DNA polymerases from distantly related organisms suggesting the need to preserve critical functions, little is known at present about their direct role in the polymerase reaction mechanism. However, some insight has derived from the analysis of Herpes virus-type 1 pol mutants with altered sensitivity to various antiviral drugs, such as pyrophosphate (PPi) and deoxyribonucleoside triphosphate (dNTP) analogs, which mimic and/or compete with the natural substrates for enzyme binding. Thus, several different mutations of Herpes polymerase involved in drug recognition were sequenced and resulted in substitution of single amino acid residues clustered in the similarity Regions 4, 5 and 8 of Fig. 3 (42). This led to the proposal that these regions fold together to form the binding site for PPi and dNTP. It is interesting to note that most of Herpes virus-type 1 DNA polymerase amino acids affected by the above mutations are conserved in the S. solfataricus enzyme and that some of them correspond to residues invariant or almost invariant in class B DNA polymerases, as shown in the sequence alignments of Fig. 3. Furthermore, it has been recently proposed (43–44) that also the sequence motif YGDTDS, almost invariant in class B DNA polymerases (similarity Region 6 in Fig. 3), is part of the dNTP binding site. Bernad et al. (43) postulated that it could play a role in metal binding at the polymerase active site, on the basis of its homology with known or putative metal-binding amino acid sequences, as originally pointed out by Argos (45). All that considered, it is tempting to speculate that the

![Fig. 3. Similarity regions among S. solfataricus and various DNA polymerases.](image-url)
carboxy-terminal portion of family B DNA polymerases defines
a structural module with polymerization function. In this context,
we consider noticeable that the 50 kDa proteolytic fragment of
*S. solfataricus* DNA polymerase, which corresponds to the
enzyme carboxy-terminal half and includes the conserved
sequence. Regions from 4 to 8 retains polymerase activity,
as previously shown by gel assay according to Karawya (20). The
same result has been recently obtained (F.M.P., unpublished
results) also using the gel assay procedure of Longley and
Mosbaug (46). By this technique, after polypeptides separation
through a SDS-gel and in situ polymerase reaction, the
radiolabeled DNA products are resolved by a second dimension
of electrophoresis through a sequencing gel. To date this is the
first report indicating, on the basis of direct functional data, that
in a class B DNA polymerase the polymerization function resides
in the enzyme carboxy-terminal half. On the other hand, the
aminio-terminal region seems to contain the 3′-5′ exonuclease
active site, as above discussed.

Taking into account all these data, the modular organization
of enzymatic activities, found in *E. coli* pol I and proposed for
the other DNA polymerases (41), seems to be conserved in class
B enzyme from the archaean *S. solfataricus*. Therefore, we
postulate that family A and B DNA polymerases evolved from
a common ancestor, as also suggested by the high evolutionary
conservation of the 3′-5′ exonuclease domain. However, the
hypothesis of evolution from independent roots cannot be
excluded, since these two classes of enzymes, in spite of the
attempts to unify their structure (37,41,46), do not appear
significantly related in the sequence of their carboxy-terminal
regions. A precise assessment of the relationships between pol
I-like and class B DNA polymerases requires X-ray
crystallography studies on a family B enzyme in order to compare
its tertiary fold with the known structure of Klenow fragment.

The high sequence similarity among *S. solfataricus* and
eukaryotic α-type DNA polymerases is in accordance with partial
protein sequence data on two short regions of the DNA
polymerase from the archaean *Pyrococcus furiosus*, recently
published (49). All these data seem to support the results of recent
studies on the evolutionary relationships of the three primary
groups of living organisms. Indeed, in a recent rooted universal
phylogenetic tree Archaea and Eucarya, sharing a common
ancestor, appear to be specific (although distant) relatives (17).
Since extremely thermophilic archaebacterial species are thought
to be the most closely related of extant organisms to the common
ancestor of all life (50), *S. solfataricus* DNA polymerase could be
similar to a primitive form of a-type replicase.

REFERENCES

257, 158–166.